Master's Thesis

Sensitivity Analyses for Informative Censoring in Time-to-Event Clinical Trials

Simon Fink

Supervisors

Prof. Dr. Kauermann
Ludwig-Maximilians-University Munich

Dr. James Bell
Boehringer Ingelheim Ltd

May 6, 2015

Abstract

The assumption of censoring at random for analyses of time to event data in the presence of informative censorings can lead to a biased estimation of the likelihood and therefore biased estimates in the cox regression.
This thesis presents three different approaches to analyse the sensitivity of time-to-event data to informative censorings. The censoring at random (CAR) method of multiple imputation via Kaplan Meier imputation (KMI) serves as a starting point for these methods. On this basis three approaches are introduced and their implementation in SAS is explained in detail. Finally, all methods are applied to a real world data set and the results are discussed.
The tipping point analysis adjusts the Kaplan Meier curve of the active treatment group used for the KMI by raising it to the power of a δ. This δ is gradually increased until the difference in the survival curve of the active treatment group and the reference group is no longer significant. The smallest value of δ which fulfills this is called the tipping point. If the resulting tipping point is not clinically reasonable, the original analysis of the data can be called robust to the CAR assumption [14]. However, it might be the case that no such δ exists, an example of which is given in the thesis.
The second approach, reference-based imputation, is based on the assumption that from the time point of their drop out onwards, patients who drop out of the active treatment group have the same risk for having an event as the patients in the reference group. Under this assumption, KMI for the active treatment group is performed using the Kaplan Meier curve of the reference group. The aim of this approach is to test if the difference in the survival curves of the active treatment group and the reference group is still significant after the imputation.
Finally, a pattern imputation approach is introduced. The implemented program allows the user to define patterns of censored patients who are believed to behave similarly after censoring. For every pattern the adjustments to the Kaplan Meier curves for the KMI introduced in the first two approaches can be applied to all treatment groups. This means every treatment can be imputed by using the Kaplan Meier curve of any treatment and a predefined δ can be applied.
Pattern imputation can be used for sensitivity analyses by making conservative assumptions for the patterns. It can also give an impression of how the data might look like without censorings if the assumptions for the pattern are based on realistic reasons.

Acknowledgements

I would like to express my gratitude to Prof. Dr. Göran Kauermann and the company Boehringer Ingelheim Pharma GmbH & Co. KG (BI) for giving me the opportunity to write this thesis in cooperation with the Department of Statistics at the Ludwig-Maximilians-University Munich and BI.

Many thanks to Prof. Dr. Göran Kauermann for taking over the supervision of my thesis. I appreciate your openness in terms of the development of the content and the constellation of Bavarian, Swabian and English collaboration.

My supervisor from BI, Dr. James Bell, deserves special thanks for providing me with this extraordinarily interesting, demanding and up-to-date topic and for giving me the chance to write my thesis at BI. Furthermore, I would like to thank you for your great effort and for always taking time to answer my questions. Your excellent supervision, your tireless rereading of my work and all your suggested improvements helped me a lot.

Furthermore, my gratitude to Dr. Hendrik Schmidt for giving me the chance to write this thesis at BI and making the connection to Dr. James Bell.
Moreover, many thanks to Julia Krzykalla for the critical rereading of my thesis.
In addition, I want to thank the whole Biostatistics group at BI for the warm welcome and the offered support whenever needed.
Finally, I would like to thank my family and my friends for the encouragement and support during the time of studying, particularly during the work on this thesis.

Contents

Chapter 1

Introduction

A Time-to-Event study examines the duration until a predefined event occurs. Originally this event mostly was death, which established the term survival analysis for this subject area. Today there are far more types of events that can be examined, for example time until people find a job after graduation or the duration of tool use until they break for example. This is the reason why the more general term ”time-to-event” is becoming more common.

If a participant in a time-to-event study leaves the study or drops out for any reason before the event of interest can be observed, the time when the drop out occurs is recorded and described as censoring.

For the analysis of the data those censorings are usually treated as non-informative or censored at random, which means they just give the information that the examined event didn't occur until the observation was censored. The assumption that the censoring is independent from the eventual event time is necessary for most of the common evaluation methods like Kaplan-Meier, Cox regression and the log-rank test.

In practise however, drop outs may be not independent from the eventual event time. In clinical trials, participants often leave a study because their health deteriorates. In this case it is likely that the event of those participants would have occurred earlier than the average event time of the other participants. In contrast, it is possible that a participant feels well enough to leave the country while the study is still ongoing. This would probably lead to an event time at a later time point than the average. Therefore, censoring may often be informative.

This informative censoring is analogous to the issue of missing not at random (MNAR) data in longitudinal clinical trials. In this case, it is assumed that a missing variable not only depends on previously measured values but also on the true measurements that could not be collected due to the drop out. Those unknown or unknowable factors affect the outcome of the respective participants of the study. Therefore the assumption of censoring at random (CAR) which would be analogous to missing at random (MAR) in longitudinal data does not hold in some cases.

For longitudinal data and if the MNAR assumption is met, several techniques have been suggested to handle missing data. One approach presented by Ratitch in 2013 [18] is to use adjusted multiple imputation. Multiple imputation uses the given information of a dataset to define a distribution for the missing variables. This is sampled multiple times and across those implemented datasets an estimate can be calculated. One form of this is a pattern mixture model, which allows formation of different patterns depending on factors like time of drop out, treatment and reason for drop out. All participants with missing variables can then be classified in a pattern which fits best to the conditions of the participant.

This approach using adjusted multiple imputation was originally designed for longitudinal data. It has recently been suggested that these ideas may also be transferable to time-to-event data [14].

In this thesis, Kaplan Meier multiple imputation, as described by Taylor et al. [30], will be used as a basis, with adjustments and/or the pattern-specific modifications created to introduce MNAR assumptions. These will be used for three approaches for a sensitivity analyses for informative censorings.

The first is a tipping point analysis which worsens the expected time of event in the treatment group step by step until the difference between the groups is no longer significant.

In the second, a reference based imputation will be used to assume the patients with censorings in the active treatment arm will not receive active treatment any more and therefore all time points for imputation are drawn from the curve of the reference group.

The third approach is a sensitivity analysis that tries to create a realistic scenario by using informations about the censorings. This will be done with a pattern mixture model that allows to from different patterns across the censorings and treat them accordingly to the informations given about them.

Chapter 2

Theoretical background

2.1 Time-to-event analyses

2.1.1 Survival time

Let T denote the time until an examined event occurs. This event can be death, the failure of a mechanic machine, the rejection of an organ after transplantation, a job offer or any other event of interest that can be observed.
More precisely T is a non-negative random variable from a homogeneous population, which can be assumed to follow a distribution[7]. Commonly, one uses one of the following three formulations to define the distribution of T mostly unambiguously: the survivor function, the probability density function and the hazard function[7]. They will be explained in more detail later in this section.
One of the characteristics of time-to-event (TTE) data is the fact that the occurrence of the event of interest is not guaranteed. It is possible, that a participant leaves the study without having an event, which would be the case if he moves to another country and can't be followed up for example. Another possibility is that the event of interest doesn't occur at all if the occurrence of the event is not guaranteed. It is also possible that the event happens but not until after the end of the study which means there is an event but it is not recorded because the study is finished by that time. This loss of information is called censoring and will be explained in the next section.

2.1.2 Censoring

Censoring occurs if a patient drops out of the study without having the event of interest or if the study ends before the examined event occurs. In that case, the exact time-to-event is not known for this participant. But the information, that the participant did not have an event until the time of the censoring still provides information for the analyses. There are different types of censorings:
Right censoring is defined as follows: Let T_i^* denote the true time-to-event and T_i the observed time-to-event.

Definition 1 (Right censoring)**.**
The time-to-event T_i is right censored, if the observed event time is bigger than a random value C_i, i.e.

$$T_i = min\{T_i^*, C_i\} \quad and$$
$$\theta_i = \begin{cases} 1 & if\, T_i = C_i, \\ 0 & if\, T_i = T_i^* \end{cases}$$

where C_i denotes the time when the variable T_i^* is censored and θ_i indicates if a censoring or an event occurred[7].

Right censorings can occur during the study, if a patient leaves without having the event of interest or if the event did not occur before the end of the study.
For the sake of completeness, it should be mentioned, that there are other types of censoring, which would be left censoring or interval censoring for example. It is also possible, that more than one type of censoring occur

in one and the same study but this thesis only deals with right censoring and so the others are ignored from this point.
For most techniques for the analysis of time-to-event data, censorings are assumed to be random or non-informative. This means that the censorings follow a distribution which is independent from the distribution of the events. Additionally it is assumed that the C_i's are independent from each other and the event times $T_1, \ldots, T_n$ [7]. This thesis however focuses on the situation of informative censoring. The influence and assumptions of this type of censoring are explained in more detail in section 2.2.

2.1.3 The Kaplan Meier estimator

The explanations given in this section are based around those given by Collett[1].

Definition 2 (Survivor function)**.**
The function

$$S(t) = P(T \geq t),$$

which is defined as the probability that the time to the examined event of a patient is bigger than some time t *is called the survivor function.*

This function can also be expressed as

$$S(t) = 1 - F(t),$$

where $F(t)$ denotes the cumulative distribution function of T.
To be able to estimate this survivor function, the Kaplan Meier estimator is used. For the calculation of this estimator the observed event times need to be sorted ascendingly, i.e. the realisations of the event times $T_i, i = 1, \ldots, n$ must fulfil the inequality $t_{(1)} \leq t_{(2)} \leq \ldots \leq t_{(n)}$. Although statistically not possible, due to the discrete recording of the time-to-event, it may happen that in practice multiple censorings and events are recorded at the same time point which is for example in days.
Let n be the total number of observations. The number of different event times is denoted by r, where $r \leq n$ for the reason, that in right censored data not all patients have events and two or more events can be observed simultaneously. Consequently r different survival times $t_{(1)} < t_{(2)} \ldots < t_{(r)}$ are involved in the calculation of the Kaplan Meier estimator. The amount of patients under risk just before $t_{(j)}$ is denoted by n_j. This value includes those patients which an event at time $t_{(j)}$. The number of patients with an event at $t_{(j)}$ is defined as d_j.

Definition 3 (Kaplan Meier estimator of the survivor function)**.**
Let $t_{(1)} < t_{(2)} \ldots < t_{(r)}$ *be sorted event times and* n_j *denotes the number of patients under risk just before the* j*-th event time* $t_{(j)}$*. Furthermore the number of patients with an event at* $t_{(j)}$ *is defined as* d_j*. The product*

$$\hat{S}(t_{(k)}) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right)$$

is then called the Kaplan Meier estimator for the survivor function.

It is an estimator for the probability that a patient has no event until the time $t_{(k)}$

2.1.4 The Cox proportional hazards model

The Cox proportional hazards model, or Cox model, is the most common method used to examine the relationship between the time-to-event and the covariates.
It is a semi-parametric model for the hazard rate.

Definition 4 (Hazard function)**.**
Let

$$S(t) = P(T \geq t)$$

be the survivor function and $\lambda(t)$ *the hazard or age specific failure rate of* t *that is,*

$$\lambda(t) = \lim_{\Delta t \to 0+} \left\{ \frac{P(t \leq T < t + \Delta t | t \leq T)}{\Delta t} \right\} \quad [2].$$

Based on this definition, one can develop an estimator for the hazard function.
With the definition of the survivor function and the probability theory of conditional events, the numerator on the formula of the hazard function can be expressed as

$$\frac{P(t \leq T < t + \Delta t)}{P(T \geq t)} = \frac{F(t + \Delta t) - F(t)}{S(t)},$$

which leads to

$$\lambda(t) = \lim_{\Delta t \to 0+} \left\{ \frac{F(t + \Delta t) - F(t)}{\Delta t} \right\} \frac{1}{S(t)} \quad [1]$$

As the definition of the derivative of F(t), denoted by f(t) is given by

$$\lim_{\Delta t \to 0+} \left\{ \frac{F(t + \Delta t) - F(t)}{\Delta t} \right\}$$

the hazard rate can be expressed as the quotient

$$\lambda(t) = \frac{f(t)}{S(t)}$$

Using the fact that $f(t) = -\frac{\partial}{\partial t} S(t)$, the following relation between the survivor function and the hazard rate can be established

$$\lambda(t) = -\frac{\partial}{\partial t}(log\ S(t)),$$

and therefore

$$S(t) = exp\{-H(t)\} \quad [1]$$

with $H(t) = \int \lambda(t) dt$ defining the cumulative hazard function.
The Cox model is then defined as follows

Definition 5 (Cox model)**.**
Let $z_j = (z_{1j}, \ldots, z_{pj})$ *be the vector of the* p *covariates for the* j*th individual. The Cox model is defined by*

$$\lambda(t, z) = \lambda_0(t) e^{z\beta}$$

where β *is a* $p \times 1$ *vector of unknown parameters and* $\lambda_0(t)$ *is an unknown function for the baseline hazard [2].*

The baseline hazard gives "the hazard function for the standard set of conditions $z = 0$"[2].

Proportional hazards assumption and hazard ratio

The proportional hazards assumption allows the elimination of the unknown baseline hazard in the formula of the hazard ratio and therefore provides an expression that is independent from time t.

Definition 6 (Hazard ratio)**.**
The ratio of hazards for an individual with covariate vector z^* compared to an individual with covariate vector z is given by

$$\begin{aligned} HR(z*, z) &= \frac{\lambda_0(t)e^{(z^*\beta)}}{\lambda_0(t)e^{(z\beta)}} \\ &= \frac{e^{(z^*\beta)}}{e^{(z\beta)}} = e^{[(z^*-z)\beta]} \\ &= \textit{constant} \quad [2][5] \end{aligned}$$

This assumption can be used to derive an estimator for the vector of unknown parameters β.
As before, let $t_{(1)} < \ldots < t_{(r)}$ be sorted times until the event. Furthermore let $R(t_{(i)})$ denote the set of individuals under risk at time $t_{(i)}$, which consists of all individuals that did not have an event or censoring up to time $t_{(i)}$.
Then, the probability of an event for individual j at time $t_{(i)}$ and d_i the number of patients having an event at time point $t_{(i)}$. Assuming only one event can occur at time $t_{(i)}$ is

$$\begin{aligned} &P(\text{subject } j \text{ fails at } t_i | R(t_{(i)}), d_i = 1) \\ &\quad = \frac{P(\text{subject } j \text{ fails at } t_i | R(t_{(i)}))}{P(d_i = 1 | R(t_{(i)}))} = \frac{P(t_j = t_{(i)} | R(t_{(i)}))}{P(d_i = 1 | R(t_{(i)}))} \\ &\quad = \frac{\lambda_0(t)e^{z_j\beta}}{\sum\limits_{l \in R(t_{(i)})} \lambda_0(t)e^{z_l\beta}} = \frac{e^{z_j\beta}}{\sum\limits_{l \in R(t_{(i)})} e^{z_l\beta}} \quad [2][5] \end{aligned}$$

Assuming the event times of different patients are independent from each other, a partial likelihood for β can be calculated by multiplying the individual likelihoods

$$L(\beta) = \prod_{i=1}^{r} \frac{e^{z_i\beta}}{\sum\limits_{l \in R(t_{(i)})} e^{z_l\beta}} \quad [5]$$

The log partial likelihood is

$$l(\beta) = \sum_{i=1}^{r} z_i\beta - \sum_{i=1}^{r} log \left[\sum_{l \in R(t_{(i)})} e^{z_l\beta} \right] \quad [2].$$

The estimator for β can now be calculated by finding the maximum of $l(\beta)$. An iterative algorithm which solves this problem is the Newton-Raphson Algorithm for example [1].
The Cox model gives estimates for the influence of the treatment and other covariates. This means when comparing the treatment effect a p-value smaller than a predefined significance level indicates a significant difference between the two treatment groups. The advantage of the Cox model is the possibility to estimate the hazard rate and under the assumption of proportional hazards allows the comparison of two treatment groups. The downside however is that the assumption of proportional hazards over time is a strong assumption that

needs to be fulfilled if the Cox model is used.
The log-rank test is another option for the comparison of treatment effects.

2.1.5 The log-rank test

The log-rank test can be used to compare two groups with each other and examine if there are significant differences in the group-specific survivor functions, i.e. $H_0 : S_1(t) = S_2(t)$ where $S_i(t)$ denotes the survivor function of group $i,\ i \in \{1, 2\}$.
Suppose there are r events $t_{(1)} < \ldots < t_{(r)}$ happening across patients in both groups. The amount of events in Group 1 will be denoted as d_{1j} and in Group 2 as d_{2j}. Furthermore it is assumed that there are n_{1j} and n_{2j} patients at risk at time $t_{(j)}$ in Group 1 and Group 2, respectively [1]. This sums up to $d_j = d_{1j} + d_{2j}$ events over both groups observed in $n_j = n_{1j} + n_{2j}$ patients at risk at time $t_{(j)}$.
The log-rank test compares the expected amount of events in one group under the null hypothesis H_0 with the observed true value at each event time.
The number of events at time $t_{(j)}$ in Group 1, can be regarded as a random variable D_{1j} with possible values between 0 and the minimum of d_j and n_{1j}. The distribution is called *hypergeometric distribution* which defines the probability that this random variable D_{1j} is equal to a certain value d_{1j} is given by

$$\frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}} \quad [1].$$

The mean of the hypergeometric distribution is given by

$$e_{1j} = n_{1j} d_j / n_j \quad [1],$$

so e_{1j} is the expected number of individuals with an event at time $t_{(j)}$ in Group 1.
Under the null hypothesis the sum of differences $d_{1j} - e_{1j}$ is expected to be zero. The resulting statistic is given by

$$U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j}) \quad [1].$$

The mean of this statistic is zero because of $E(d_{1j}) = e_{1j}$. Knowing that d_{1j} is hypergeometricly distributed, the variance for one single event time $t_{(j)}$ is given by

$$v_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \quad [1]$$

which leads to

$$var(U_L) = \sum_{j=1}^{r} v_{1j} = V_L \quad [1]$$

as the event times are independent from each other.
It can be shown, that with a sufficient number of events, U_L is approximately normally distributed implying that $U_L / \sqrt{V_L}$ follows a standard normal distribution

$$\frac{U_L}{\sqrt{V_L}} \sim N(0, 1) \quad [1].$$

Furthermore the square of the quotient is chi-squared distributed with one degree of freedom

$$\frac{U_L^2}{V_L} \sim \chi_1^2 \quad [1].$$

With the knowledge of the distributions a p-value can be calculated accordingly.
The log-rank test is a powerful tool to test if there are differences between two treatment groups but the only information it provides is if the difference is significant or not. It is not possible to derive a factor from the statistic which says patients in group A have x-times later event times than patients in group B.
A different derivation of the log-rank test is given by Peto [16].

2.2 Missing data assumptions

In longitudinal data it happens that values are missing for one or more time points for a patient. The techniques to analyse the data require different assumptions for the missing data. If a complete case analysis (whereby subjects with relevant missing data are ignored) is done the missings have to be missing completely at random (MCAR) to get a valid result. Likelihood-based methods and Bayesian methods require only the less strict assumption of missing at random (MAR).
Besides this assumption, there is one more type of missingness: missing not at random (MNAR)[25].
The relation between those assumptions and the data can be described as follows. Let Y_{com} be the complete data set, which consists by the observed (Y_{obs}) and the missing (Y_{mis}) data. The event of missingness is represented by the random variable M.
MCAR assumes that the missing is neither dependent on the values observed before the missing (including covariates) nor on the data that would have been measured, if the missing would not exist,

$$P(M|Y_{com}) = P(M).$$

MAR does not depend on the missing data, but on the observed observations. For the probability of missingness this means,

$$P(M|Y_{com}) = P(M|Y_{obs}, X).$$

If the distribution depends on both the observed data and the missing data it is called missing not at random,

$$P(M|Y_{com}, X) = P(M|Y_{obs}, Y_{mis}, X) \qquad [25].$$

In time-to-event data missing values generally are referred to as censorings. Since for most analysis techniques the censorings are assumed to be independent from the observed values aka events but not from the covariates the censoring at random (CAR) assumption can be compared to the MAR assumption in longitudinal data [6]. This means, that the participants who drop out of the study don't provide any further information about the event time other than they didn't have an event until the censoring occurred. Differently spoken, if a subject drops out of a clinical trial before progression, then it is assumed that, he may have progressed like the average of the study population.
In practice, however, the censorings sometimes do give informations, which would be called informative censoring.Those informative censorings are similar to the MNAR assumption in longitudinal data since the censorings are not independent from the observed events and the covariates may influence the time of the censoring. This means, that a patient who leaves the study drops out for a reason directly or indirectly related to the event

of interest. This is for example the case if a patient discontinues from a clinical trial because they feel that the drug isn't effective, so they probably would be more likely to experience a (negative) event sooner. Deleting those cases can bias the outcome of a study. If the benefit of the drug over the reference group is just slightly significant, when excluding those patients , the superiority may be no longer significant.
As mentioned most analysing techniques assume CAR. The importance of this assumption can be seen in the calculation of the likelihood where "the only contribution to the likelihood is just the probability that lifetime T exceeds" the censoring time [28]. Thus the assumption of CAR in the presence of informative censoring would bias the likelihood and therefore deliver wrong estimates for β.
One way to deal with informative censoring is imputing the censored values under predefined assumptions about their behaviour after the censorings.

2.3 Imputation

The explanations in this section are based around those given by Little and Rubin's "Statistical Analysis with missing data" [10]

2.3.1 Single Imputation in Longitudinal Data

Most analysis methods in statistics require complete data, that is values are available for each measuring time point, each variable and each patient. In practice however, this is rarely the case, i.e. there are missing values in for longitudinal data or censorings in TTE data respectively. The imputation of those missings or censorings is one possibility to deal with them.
The term imputation describes the process of filling in reasonable values for the missing values and create complete data [13]. Up to now, imputation in TTE data is not very widespread but it is used in longitudinal data commonly. Therefore, in this thesis, imputation methods are described in the framework of longitudinal data and the extension to TTE data is explained subsequently.
The idea of imputing missing data is relatively straightforward, however there are many ways of doing it and several difficulties that have to be overcome. The next step is to get different options for its realisation.

Unconditional Means

A simple approach for imputing missing values is the imputing of unconditional means. Let $Y = Y_1, \ldots, Y_p$ define the matrix of covariates where each column contains one covariate. If $Y_j = (Y_{1j}, \ldots, Y_{nj})$ denotes the vector of measurements for variable j and $\bar{Y_j}^{(j)}$ is the mean of all these values then all values Y_{ij} that are missing are imputed with the mean $\bar{Y_j}^{(j)}$.
This technique preserves the mean of Y_j but the variance, which would be $v_j^{(j)}$ under MCAR and a consistent estimate of the true variance is $[(n^{(j)} - 1)/(n - 1)]v_j^{(j)}$, where $n^{(j)}$ denotes the number of observed cases and n the overall number of patients.
This means that by imputing unconditional means, the variance is underestimated by a factor of $(n^{(j)} - 1)/(n - 1)$. Also the shape of the empirical distribution is different after the imputation, which can lead to misunderstandings when studying the data with histograms or other plots.

Last observation carried forward

Last observation carried forward, or LOCF, is applied when no measures follow after a missing value for a patient. More precisely: if in a study with time points $1, \ldots, p$ patient i has measures $Y_{i,1}, \ldots, Y_{i,q}$ with $q < p$ the last observation $Y_{i,q}$ is used to impute all missing values $Y_{i,q+1}, \ldots, Y_{i,p}$.

Hot Deck Imputation

According to this method, missings are imputed by random sample from the empirical distribution which can be estimated from the underlying data set. Usually one uses the empirical distribution of the observed values of the variable only.
If the patients are assumed to show different outcomes depending on an underlying structure of clusters (group/-stratum), it is possible to calculate a separate empirical distribution function for each group or stratum in the study and for each missing value draw a random value from the corresponding distribution. On average, the means remain equal to those of the original data set.

Regression Imputation

If for an individual in a longitudinal data set some values have been observed and some are missing it is possible to create a model which estimates the missing values using the information given by the observed ones. The regression model can include information which may not be contained in the distribution of the variable, such as the correlation with other variables.

The presented imputation techniques are a choice of examples for longitudinal data. The same ideas can be translated for the application to TTE data to impute censorings and get a complete data set with 100% event rate and no censorings in theory. In practice, as an extrapolation beyond the end of study date can lead to serious bias, it is only possible to impute the values for those patients that drop out before the end of the study but not the ones that are right censored at the end of the study.

2.3.2 Single Imputation in Time-to-event Data

In longitudinal data, missing values do not provide further information. A censoring in TTE data however gives the information, that the patient has had no event until the time of censoring. So for the imputation of a censoring, it is important that this is also taken into consideration, i.e. that the imputed event has to be later than the censoring.
Let $T_1^*, \ldots, T_n^*$ denote the actual event times of the patients disregarding any potential censorings until the event happens and $C_1, \ldots, C_n$ denote the potential censoring times. The value observed in the study will then be $T_i = \min(T_i^*, C_i)$ which means that if the actual event time is smaller than the potential censoring, the time of the event is captured in the study data but if the censoring has the smaller valueonly the censoring time will be captured for this patient because the real event time T_i^* is not known[30].
Let $n_{mis} = n - n_{obs}$ denote the number of censorings in the study where n_{obs} is the number of observed events. The number of patients under risk just after t_j is defined by the risk set $R(t_j) := \{i : T_i > t_j, i = 1, \ldots, n\}$.
For imputing censorings equivalent techniques as for the longitudinal data can be used. The mean life time could be estimated and inserted for each censoring, which would be an equivalent to the unconditional means imputation. The methods explained in the following for TTE data are related to the hot deck imputation. The basis for both of the examples is to calculate a distribution derived from the current risk set for each censoring [30].

Risk set imputation

The data set contains n_{mis} censored times t_j that should be imputed. For each of these missings, a pair (t_i^+, δ_i) is randomly drawn from the risk set $R(t_j)$ of $t_{(j)}$. All pairs (t_i^+, δ_i) in the risk set are drawn with equal probability, the event times as well as the censored times. A censoring that has been imputed again with a

censoring $(t_i^+, 1)$ will not be imputed a second time but stay a censoring at a later time. The censorings will be imputed ascendantly, beginning with the one that occurred first in the study. If the last observation is a censoring it will not be imputed due to the reason that there are no pairs of observations in the risk set [30].

Kaplan Meier imputation

Kaplan Meier imputation (KMI) calculates a separate survival curve for every censoring and imputes the censored value by drawing a random value from the survivor function.
Analogously to the risk set imputation, the procedure begins with the smallest censoring time and the censorings are imputed ascendantly. For each censoring time c_k, an individual survival curve $\hat{S}(c_k)$ can be calculated on the basis if the corresponding risk set $R(t_k)$[30].
Therefore, it is assumed that the study includes M observed events at distinct times $(t_1 < t_2 < \ldots < t_M)$ and K distinct censoring times $(c_1 < c_2 < \ldots < c_M)$. It is also assumed that there may be more than one patient with the same times at risk y_i, which means ties are allowed.
For calculating the individual survivor functions, let k index the censoring times and $t_{k,0}$ denote the latest failure time prior to c_k (or equal to it) when $t_1 \leq c_k$. If $t_1 > c_k$ then let $t_{k,0} = 0$. The jth failure time after c_k is denoted $t_{k,j},\ j = 1, \ldots, J_k$, when $c_k < t_k$. From the data (Y_i, θ_i) the Kaplan-Meier estimates $\hat{S}(t_{(k)})$ fot the survivor distribution can be calculated. For a censoring time c_k followed by at least one failure time, the estimate of the survivor function $\hat{S}(c_k)$ is defined by linear interpolation as follows:

$$\begin{aligned}\hat{S}(c_k) &= \hat{S}(t_{k,0}) - \frac{c_k - t_{k,0}}{t_{k,1} - t_{k,0}} * (\hat{S}(t_{k,0}) - \hat{S}(t_{k,1})) \\ &= \frac{(t_{k,1} - c_k)\hat{S}(t_{k,0}) + (c_k + t_{k,0})\hat{S}(t_{k,1})}{(t_{k,1} - t_{k,0})} \qquad [35]\end{aligned}$$

The survival curve gives the probability of not having an event up to a certain time and therefore takes values in the interval $[0, 1]$. This characteristic is used to find an imputation time by generating a realisation of a uniform $(0, 1)$ value and reading the corresponding t_i^+ from the estimated survivor function [30] [35].
This procedure is valid as long as the resulting t_i^+ is smaller than the largest event time. However, it is also possible to produce a value beyond the largest event time. In this case the censoring is imputed as a censoring at the last observed event time. Censorings that happen after the last event are not imputed because there are no events left to construct a survival curve. The imputed values are not used to calculate the survival curves for the subsequent censoring times. They are saved in a separate data set and used to replace the corresponding censorings after the complete imputation process is completed [30]. Finally, a TTE analysis can be done as planned before imputation.

In practise imputations in TTE data are not very common. This is due to the reason that until recently it has been believed that there are no real benefits for using it: The original data can be analysed with methods valid for CAR and this also applies for the imputed data.

2.3.3 Delta adjustment

Zhao et al. [35] proposed an extension to the Kaplan Meier imputation introduced in Taylor et al. [30]. In this, a fixed hazard ratio δ for a patient who drops out of the study before having an event is introduced relative to the patients still remaining on their assigned treatment is introduced as the sensitivity parameter. The estimated survivor function at time t is equals $\hat{S}(t)^\delta$ under the proportional hazards assumption. For a

patient with censoring at time $c_k < t_M$, the estimated conditional probability of having an event in the time interval $[t_{k,j}, t_{k,j+1}]$ for $j = 1, \ldots, (J_k - 1)$, is given by

$$\hat{f}_{k,j}(\delta) = \frac{\hat{S}(t_{k,j})^\delta - \hat{S}(t_{k,j+1})^\delta}{\hat{S}(c_k)^\delta} \qquad [35]$$

After defining this probability the conditional cumulative incidence function for a censoring at time c_k to have the event by the time t in $[t_{k,j}, t_{k,j+1}]$ can be calculated by summation of the $\hat{f}_{k,j}(\delta)$ for the respective time intervals:

$$\hat{F}_{k,j}(\delta) = \sum_{r=0}^{j} \hat{f}_{k,r} = 1 - \frac{\hat{S}(t_{k,j+1})^\delta}{\hat{S}(c_k)^\delta} \qquad [35]$$

With this formulation, $\delta > 1$ (or < 1) implies a higher (or lower) hazard after c_k for patients censored at that time point compared to patients still under risk at c_k.
For the reference group the δ after the drop out is considered to be lower compared to patients who disband from the active treatment group. Assuming $\delta_p = 1$ would imply a CAR like behaviour after the drop out for the reference group. This way the hazard in the active treatment group becomes the only parameter for calibrating the sensitivity analyses $\delta_T = \delta$ [35]. The realization of this delta adjustment is done by using the conditional incidence function instead of the survivor function $\hat{S}(c_k)$ in the Kaplan Meier imputation.

2.3.4 Multiple Imputation

Since the single imputation methods cannot take into consideration the uncertainty in the data and for some imputations adjustments are anyway necessary to get appropriate results, the multiple imputation approach is one possibility to deal with this [10][17][22][23][24].
Multiple imputation is done similar for longitudinal and TTE data. The idea is to impute a data set M times using a single imputation technique that uses a random distribution and to combine the calculated estimates from the imputed data sets [30]. This combination is usually done by means of Rubin's rules which require normal distribution of the estimates.
Let $\theta_l, V_l, l = 1, \ldots, M$ be the M estimators and variances resulting from the single imputations, i.e. one pair (θ_l, V_l) denotes the estimate of interest and its variance. In longitudinal data this could be a point estimate and for TTE data this could be the log hazard ratio. The combined estimate for $\theta_1, \ldots, \theta_M$ is the mean of all estimates of the single imputations,

$$\bar{\theta} = \sum_{l=1}^{M} \frac{\hat{\theta}_l}{M} \quad [10][30].$$

For the variance, two components have to be included: the average within-imputation variance,

$$\bar{V} = \sum_{l=1}^{M} \frac{\hat{V}_l}{M} \quad [10][30],$$

and the between-imputation variance

$$B = \frac{\sum_{l=1}^{M} (\hat{\theta}_l - \bar{\theta})^2}{M - 1} \quad [10][30].$$

In case of TTE data the variances $\hat{V}_l$ are calculated using the Greenwood formula [4].
If θ is a vector, $(\cdot)^2$ is replaced by $(\cdot)^T(\cdot)$. The total variance of $\bar{\theta}$ is then given by the sum

$$T = \bar{V} + \frac{M+1}{M}B \quad [10][30].$$

If θ is a scalar the distribution to be used for interval estimates and significance tests is a t-distribution,

$$(\theta - \bar{\theta})T^{-1/2} \sim t_v \quad [10][30],$$

where v denote the degrees of freedom

$$v = (M-1)\left[1 + \frac{1}{M+1}\frac{\bar{V}}{B}\right]^2 \quad [10][30].$$

As the singly imputed values are drawn from the distribution of the observed values or events when imputing TTE data, the mean of the imputed values is similar to the mean of the observed ones as long as no covariates are included in the imputation process. Hence, the mean of the estimates $\bar{\theta}$ does also not change after imputation under the assumption of CAR [10].
Multiple imputation has the positive characteristics of the single imputation which means the same estimates are produced and attempts to correct for the disadvantages especially the underestimation of the variance by repeating the single imputation step multiple times [10].

2.4 Pattern mixture models

As mentioned before, the reason for drop out can be independent of the response and treatment or linked to them in some way. A pattern mixture model incorporates patterns for groups of patients that are believed to behave similarly after a drop out. Therefore M represents an indicator vector where the i-th element is 0 if an actual event is observed and 1 if it is a censoring. Y denotes the matrix of outcome data composed of the missing Y_{miss} and the observed data Y_{obs} [8][9].
A pattern mixture model allows to express the joint probability of Y and M by means of conditional probabilities:

$$P(Y_{obs}, Y_{mis}, M|X) = P(M|X)P(Y_{obs}, Y_{mis}|M, X) \qquad [9].$$

In this formula $P(M|X)$ represents the "conditional probability distribution of missingness"[18] based on the matrix of observed covariates. The probability distribution of Y conditioned on M and X is denoted by $P(Y_{obs}, Y_{mis}|M, X)$.
The problem is, that some parts of $P(Y_{obs}, Y_{mis}|M, X)$ can not be estimated. Explicitly it is not possible to derive a relationship between Y_{mis} and X, due to the non-existence of Y_{mis}[18].
There are some approaches in literature which deal with this issue. One of them splits up the probability distribution as follows:

$$\begin{aligned} P(Y_{obs}, Y_{mis}, M|X) &= P(M|X)P(Y_{obs}, Y_{mis}|M, X) \\ &= P(M|X)P(Y_{obs}|M, X)P(Y_{mis}|Y_{obs}, M, X) \quad [18] \end{aligned}$$

The term $P(Y_{obs}|M, X)$ is a model for the observed data conditioned on the patterns of missingness M and the matrix of explanatory variables X, whereas $P(Y_{mis}|Y_{obs}, M, X)$ presents the model for the missing data depending on M, X and the observed data Y_{obs}.

A pattern mixture model can be implemented as follows[13][18]:

- Choose patterns with similar characteristics or reasons for missingness/censoring
- Specify a link between the distribution of unobserved/censored data and the distribution of data where the corresponding data items are observed
- Estimate one or more models on the basis of the observed data

2.5 Bootstrap

The main goal of bootstrapping is to estimate the distribution, variance or bias of a random variable $T(X, F)$, which depends on the random sample $X = (X_1, \ldots, X_n)$ and an unknown distribution F. It is assumed that the random sample follows the distribution F,

$$X_i = x_i, \quad X_i \overset{i.i.d.}{\sim} F \qquad i = 1, \ldots, n$$

where $x = (x_1, \ldots, x_n)$ is a realisation of X [3].
Hence bootstrap is used if the sample size is small and it's difficult to make a statement about eventual asymptotic behaviour or if no parametrical assumptions should be met.
The bootstrap is based on repeatedly resampling the observed data and estimate the parameters of interest from those "bootstrap samples"[3].
The principle of the bootstrap method is as follows:

- Construct the sample probability distribution $\hat{F}$, with equal probability mass $1/n$ for each data point $x_1, x_2, \ldots, x_n$.
- Draw a random sample of realizations of size n from $\hat{F}$ and denote them by

$$X_i^* = x_i^*, X_i^* \overset{i.i.d.}{\sim} \hat{F} \qquad i = 1, \ldots, n.$$

- Call this the bootstrap sample $X^* = (X_1^*, \ldots, X_n^*)$ and $x^* = (x_1^*, \ldots, x_n^*)$ its realisation [3].
  Since the sample is selected *with* replacement, observations from the original data can be included once, multiple times or not at all.
- With this bootstrap sample at hand, the measures of interest can be estimated by means of an approximate statistic $T^*(X^*, \hat{F})$

$$x^* = (x_1^*, \ldots, x_n^*) \rightarrow T^*(X^*, \hat{F})$$

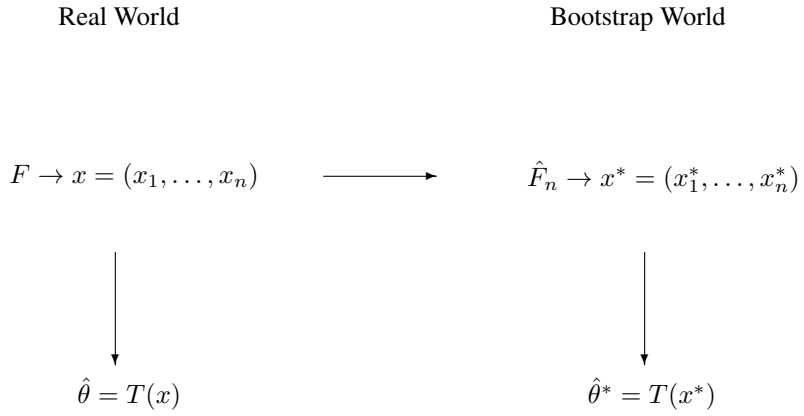
All of these steps are repeated B times, leading to B bootstrap samples $x^{*1}, \ldots, x^{*B}$ and B estimates for the value of interest $T^{*1}, \ldots, T^{*B}$[3].
With those statistics the characteristics of the distribution of T can be estimated. For example the variance

$$Var_F(T) \approx \widehat{Var}_{Boot}(T) = \frac{1}{B-1} \sum_{b=1}^{B} [T(x^{*b}) - \bar{T}_{Boot}]^2$$

or the mean of T

$$\bar{T}_{Boot} = \frac{1}{B} \sum_{b=1}^{B} T^{*b} \qquad [3].$$

The connection between the so called "bootstrap world" and the real world with the observed data is described by the following graph



The unknown distribution F gives the observed data x, which can be used to calculate the empirical distribution function $\hat{F}_n$. The empirical distribution function in its turn gives the bootstrap samples by means of resampling. The realisation of the statistic $T(x)$ is part of the real world and can be estimated from the observed data x, whereas the bootstrap replication of $T(x), T(x^*)$ is calculated using the bootstrap sample.
This method is known to give a better estimate for the variance of a statistic, especially if the sample size is small.

Chapter 3

Macros for censoring at random

The assumption of censoring at random is made in most TTE analyses. This means that the events and censoring happen independently from each other and are independent of what has been observed or will be observed. Hence, the outcome will remain unaffected.
Using a multiple imputation method on random censored observations to replace them by events should lead to the similar results as the original not imputed data. For this reason a multiple imputation is not that useful since it requires more work for the same outcome and introduces a small amount of additional variance (dependent upon the number of imputations) from the multiple imputation technique itself.
But a macro imputing random censorings can also be used as a basis for a program to deal with informative censorings. Therefore the following macro which imputes under a censoring at random assumption is presented as a starting point for the subsequent ones applicable for the case of informative censorings.

3.1 Data

For the implementation and the results of the different imputation approaches presented later the LUX-Lung 3 oncology study serves as a practical example. This section introduces the example data set, explains the structure needed by the macro to work correctly and shows examples of the results.

Example study

To introduce the data set an excerpt from the publication of the study is given here:
"LUX-Lung 3 was a global, randomized, open-label phase III study comparing first-line afatinib with cisplatin plus pemetrexed chemotherapy in patients with advanced lung adenocarcinoma and proven EGFR [epidermal growth factor receptor] *mutations. The primary end point was PFS* [progression-free survival]*, defined as time from random assignment to progression (as determined by independent blinded review) or death"*[27].
The study included 345 patients with advanced lung adenocarcinoma and confirmed *EGFR* mutations. Those were randomly assigned to treatment in a two-to-one fashion which leads to 230 patients in the treatment group and 115 patients in the reference group[27]. The overall censoring rate is 28.7% which corresponds to a total number of 99 censorings divided into $54(23.48\%)$ in the treatment group and $45(39.13\%)$ in the reference group.
The analyses were done via a stratified log-rank test and Cox proportional hazards models. The stratification factors for the log-rank test are EGFR and race (Asian or non-Asian). The Cox model included the same stratification factors and additional model covariates such as sex, age and smoking status. The results presented in this thesis were calculated from a data set containing only EGFR and race as covariates.

3.1.1 Structure needed for the macros

Most characteristics of the input data set can be directly detected and appropriately handled by the macros. Only minimal information needs to be handed over in the call of the main macro, e.g. the names of the data set and the treatment and censoring names variables don't need to be of a certain form but are to be specified in the corresponding argument. Also the stratification factor names need to be entered in form of a list for the

"Stratum" argument for the analysis later in the program.

```
%main(  data=dataname,
        Separate="YES/NO",
        Analyse="YES/NO",
        Stratum=Covariates,
        Treatment=treatment,
        Ref=Value for reference group (1,2,3,...),
        censor=censoring,
        btloop=B,
        seed=1111,
        bt="YES/NO",
        test="HR/LR");
```

The other arguments define what the program is doing during its execution. The "seed number" represents a fixed number for all random procedures in the macros to offer the possibility of reproducing the results. The *"separate="* statement steers the option to do the imputation separated only by treatment ($separate = NO$), or by all stratum groups ($separate = YES$) given in the "Stratum" argument. The same applies to the analyse statement: choosing YES means all covariates are included in the model for the analysis, using $analyse = NO$ forces the program to only use the treatment effect in the model. The value for the reference group can be any number in the range of the amount of treatments, it is not allowed to be a letter in this program, the $Ref =$ statement allows the user to specify the code of the reference group. The $btloop =$ line gives the number of data sets that should be imputed and the *bt* statement enables the user to decide whether a bootstrap is wanted or not. The last command asks if the analysis should be done by Cox proportional hazards model or by a (stratified) log rank test.

The program is structured in 5 macros with different objectives. The $\%main(\cdot)$ statement calls the macro with the same name. The *Adjustment* macro finds its use in chapter 4 and will be explained in the mentioned chapter. The other three macros, *curves*, *imputation* and *analysis* are introduced in the course of the next section containing the explanation of the whole program.
The next sections explain how the information entered in the $\%main(\cdot)$ statement is used to do the multiple imputation with the assumption of non-informative censoring.

3.2 The program

3.2.1 Preparations

The main macro transforms the data sets to the right structure and extracts the information that is needed to do a multiple imputation.
Before the actual imputation process can be started, some preparations are necessary, beginning with the import of the data set defined in the call of the macro. The next step is to define one variable which contains the number of treatment groups as well as the number of stratum groups in the case of separate imputation. In this case, the treatment group numbers and the stratum group numbers are combined into a single, separate variable. An example for a data set with 2 treatment and stratum groups each is displayed in table 3.1.

	Treatment 1	Treatment 2
Stratum 1	1	3
Stratum 2	2	4

Table 3.1: Example of the enumeration of the *strat* variable under stratified imputation

The encoding is chosen such that all stratum groups in one treatment group are imputed one after the other before the loop for the next treatment begins. If no stratified imputation is planned the variable *strat* is instead set equal to the treatment variable but it is defined either way so the program can work with the same variable regardless of the *Separate*= statement. This is a useful detail in the next chapter where the imputation procedure is modified for some of the treatment groups.
The newly defined *strat* variable can then be used by the program to determine the number of different groups that need to be imputed, each of them according to a separate Kaplan Meier curve. For each of the groups, a separate data set is created which only contains the censorings within this group. These censored data sets are used later on to replace the censoring times with the imputed event times. Taking up the previous example of 2 treatment and stratum groups, figure 3.1 shows the creation of the censored data sets.

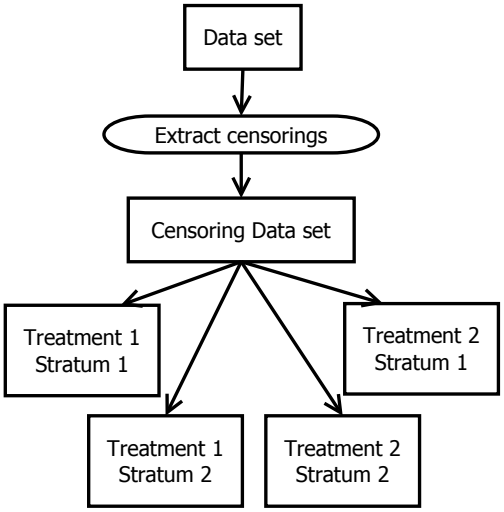


Figure 3.1: Example of the creation of the censored data sets in the case of 2 treatments and 2 strata

These steps only need to be conducted at the beginning of the program. Once the censored data sets have been created and the information about the number of groups has been determined, the group-specific Kaplan Meier curves for the imputation can be calculated. This is done via the *curves* macro which is called by the *main* macro at this point. The remaining steps in the *main* macro are described at a later point of this thesis for the sake of better understanding.

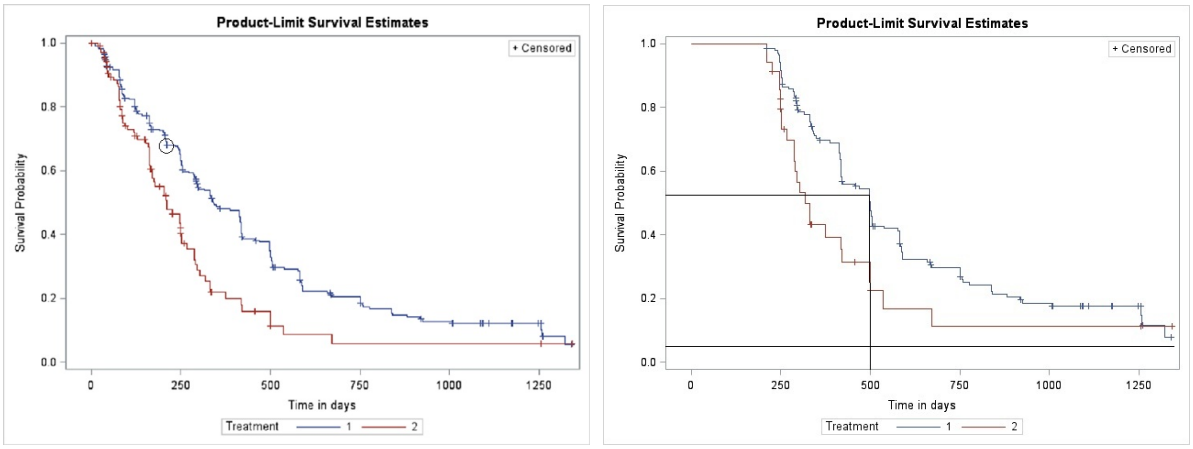
3.2.2 Calculation of the Kaplan Meier curves

The *curves* macro calculates the Kaplan Meier curves that are used in the subsequent imputation process. This can be done with or without bootstrapping.
If no bootstrap is used, the multiple imputation is done by means of the same KM curves which are calculated

per treatment group and per stratum from the originally supplied data set.
If a bootstrap is intended to be done, some preparation steps are necessary before the actual imputation can begin.
As stated before, each bootstrap sample contains a different set of observations. Therefore the Kaplan Meier tables for each data set are different. At this stage, the KM tables are calculated for each treatment (by stratum if necessary) in each bootstrap and stored in a data set. One imputation will be based on the KM curves estimated for each bootstrap sample. Hence, the number of bootstrap samples has to be equal to the number of single imputations.
The benefit of using a bootstrap is to obtain a better estimate for the variance. Without the bootstrap, the variance would be understated as the KM curves used in the imputation are derived from the sample data and not the population. This paragraph gives a short overview of the bootstrap implemented in the macro.
The macro uses a bootstrap with replacement, which draws as many observations from the original data set as are contained in the original one. This means that in one bootstrap sample, some observations are included multiple times, some only once and some are not included at all. Each data set that is going to be imputed from is build by a bootstrap and they are therefore all slightly different to each other.
Furthermore the number of events and censorings differ across the bootstrap samples and therefore they also differ in the treatment and stratum groups. The *imputation* macro imputes all censorings that occur before the last event. Since every group may have a different final event time in each bootstrap sample, the *curves* macro must also calculate the number of censorings that can by imputed in every group prior to the actual imputation.

3.2.3 Kaplan Meier imputation

Once all censorings are available in data sets, separated by their treatments (and, if required, strata) and the corresponding KM tables have been determined the actual imputation can be conducted by means of the *imputation* macro. It can impute any arbitrary data set if the Kaplan Meier curves for each group, the censoring data sets, the variable name of the time variable, the number of strata and the name of the censoring variable are entered as input parameters. As mentioned in the theoretical background 2.3.2 Kaplan Meier imputation is one method to impute a TTE data set. It uses the calculated KM tables to create a random event according to the KM curve created by the current KM table. Prior to the description of how the imputation was implemented in the macro, the KMI will be explained by means of an example.



(a) Original Data

(b) Data with $ept >= 211$

Figure 3.2: Example of KMI for a censoring at day 211

Figure 3.2a shows the Kaplan Meier curves for the original data set separated by treatment group. The censoring, which is circled in the left plot, is the one to be imputed at day 211. In a first step, all observations $<= 211$ are excluded and the Kaplan Meier curve is recalculated on the basis of the remaining observations (see figure 3.2b). Suppose the generated random uniform $[0, 1]$ value is 0.52. The corresponding upper black line intersects with the Kaplan Meier curve at day 499 which will be the imputed value. The second black line, representing another random uniform variable, does not intersect with the Kaplan Meier curve at all. In this case imputing an event is not possible so the censoring will be replaced with a censoring at the last observed event in the corresponding group.
Having this example in mind, the explanation for the implementation of the KMI in the macro is given in the following.

First the program needs to know how many censorings need to be imputed in the present *strat* group. If there are censorings later than the last event in this group, these can't be imputed because there are no following events which can be used to calculate an appropriate Kaplan Meier curve. Therefore the program identifies the last event in the *strat* group and based on that determines the last censoring that can be imputed. The censoring data sets are arranged in ascending order by the censoring times and therefore the sequential number of the last censoring that should be imputed can be used as upper limit of the imputation loop.
The imputation starts by extracting the identification number and the censoring time of the first censoring from the first censoring data set. A corresponding new KM table is created by deleting every event or censoring prior to the censoring to be imputed within the treatment/stratum group the censoring belongs to. The remaining survival probabilities are divided by the first value so the biggest value in the table is 1 again. This new calculated KM table has the same characteristics as any other KM table calculated by a Kaplan Meier estimator. In particular, the range of the risks are between $[0, 1]$ as they represent the probability of not having an event until the corresponding time. This fact is used to generate an imputation time by drawing a random uniform $[0, 1]$ value denoted by b. In SAS this can be realized by means of the *Call Ranuni* routine. According to the example above there are two possibilities of how the imputation is done:

1. If there is a value smaller than b in the KM table, the time corresponding to the largest value less than b is saved as the imputed time. Additionally, the variable censoring variable is changed from 1 to 0 which means the imputation is treated as an event in the following analysis.

2. In case b is smaller than every probability in the KM table the censoring cannot be imputed as an event because there is no corresponding time for this particular random number. In this case the latest observed event time available within the corresponding treatment/stratum group in the present bootstrap sample is saved as the imputation time for this censoring. The variable for the censoring remains 1. This means the censoring is still a censoring after the imputation but it occurs at the latest event time in the corresponding treatment/stratum group.

With the imputed time and the adjusted censoring variable, the observation can be saved in a data set that contains only imputed observations and is used to replace the censored data sets later on.
These steps need to be performed for all censorings in all treatment/stratum group with one exception. For censorings occurring later than the last event time there is no possibility to calculate a Kaplan Meier estimate due to the lack of events. Those censorings are not imputed and transferred unchanged to the imputed data set. At this stage, imputed data sets are available for every treatment/stratum censoring data set. The next subsection describes the combination of the separately imputed data sets and the subsequent analysis of the resulting data sets.

3.2.4 Analysis of a singly imputed data set

The combination and analysis of the imputed data sets is done in the *analysis* macro.
The output of the imputation process comprises separate data sets containing the imputed censored values for the specific *strat* group. Those data sets are combined in one data set containing all former censorings, most of them imputed. At this point, all events from the current data set can be extracted and combined with the data set with imputed events. The result is a singly imputed data set with the same number of observations as the original data set and comprising almost entirely events. The reason for an event rate of less than 100% in spite of the imputation is due to the fact that some censored values are later than the last event and therefore can't be imputed via KMI.

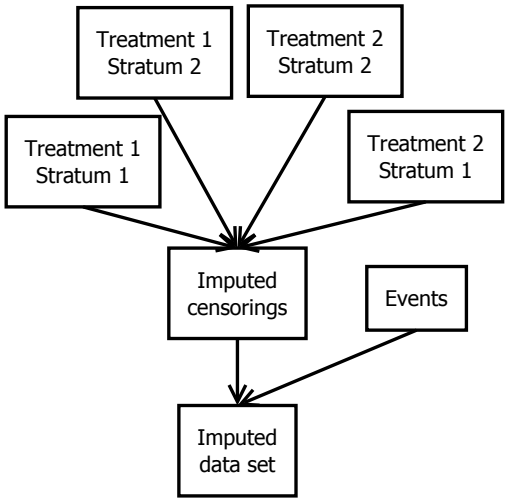


Figure 3.3: Example of the combining of the imputed data sets

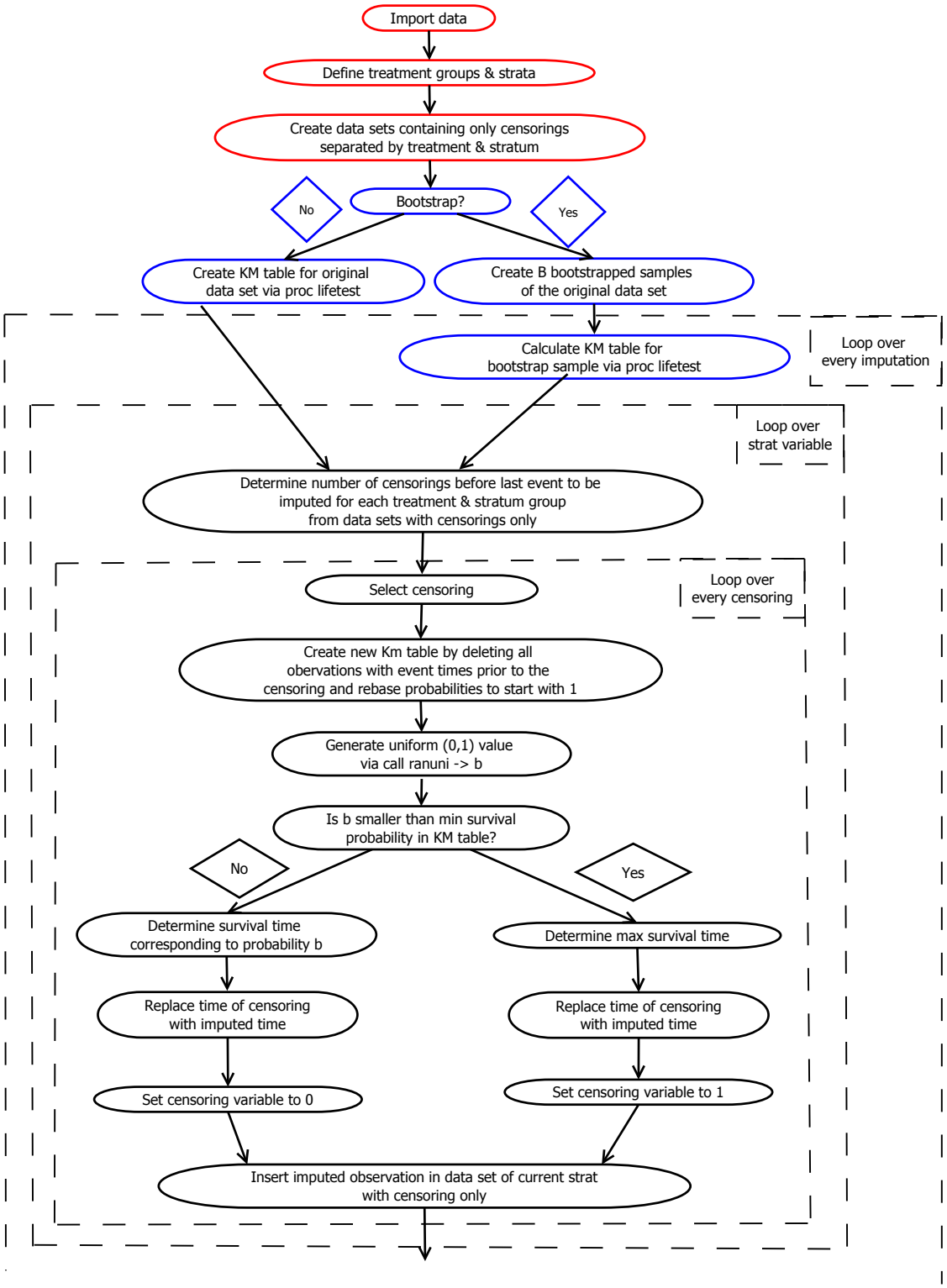
Figure 3.3 continues the previous example with 2 treatment groups and 2 stratum groups and shows a schematic of the combination of the 4 imputed data sets to the final imputed data set.
This imputed data set can then be analysed analogously to a non-imputed data set. The macro includes implementations of the Cox proportional hazards model and the (stratified) log-rank test. If the analysis is planned to be stratified the stratification factor will be included as *strata* statement in either the *lifetest* or the *phreg* procedure. For the log-rank test the treatment variable is entered as *test* statement. The Cox model includes the treatment variable as covariate. The results are saved in a data set.

The procedure described above leads to one imputed data set and the results from its analysis. As mentioned the analysis is implemented in the *analysis* macro which includes all steps that need to be done for every subsequent data set in one loop. This loop is done for each of the B data sets. All results from the analyses of the imputed data sets are saved in a data set so they can be pooled in the *main* macro.

3.2.5 Combining the results

The final steps of the program are implemented in the *main* macro again. Once all B singly imputed data sets are analysed separately by means of the above mentioned methods, the results need to be combined across imputations. For normally distributed variables, this can be done via the SAS procedure *MIANALYZE*. The distribution requirement is met for both, the log-rank statistic U_L derived in section 2.1.5 and the log-transformed hazard ratio, at least approximately [15].
The *MIANALYZE* procedure calculates the pooled mean as well as the standard error for the variables inserted in the *modeleffects* statement. It also calculates a 95% confidence limit and the p-value for the hypothesis of the variable being different to 0.

After the detailed explanations of the single steps implemented for a multiple imputation procedure, figure 3.4 provides an overview over the structure of the whole program.
The flowchart reveals that the program is straightforward and only branches to offer more flexibility for the imputation.


Import data
Define treatment groups & strata
Create data sets containing only censorings separated by treatment & stratum
Bootstrap?
No
Yes
Create KM table for original data set via proc lifetest
Create B bootstrapped samples of the original data set
Calculate KM table for bootstrap sample via proc lifetest
Loop over every imputation
Loop over strat variable
Determine number of censorings before last event to be imputed for each treatment & stratum group from data sets with censorings only
Select censoring
Loop over every censoring
Create new Km table by deleting all obervations with event times prior to the censoring and rebase probabilities to start with 1
Generate uniform (0,1) value via call ranuni -> b
Is b smaller than min survival probability in KM table?
No
Yes
Determine survival time corresponding to probability b
Determine max survival time
Replace time of censoring with imputed time
Replace time of censoring with imputed time
Set censoring variable to 0
Set censoring variable to 1
Insert imputed observation in data set of current strat with censoring only

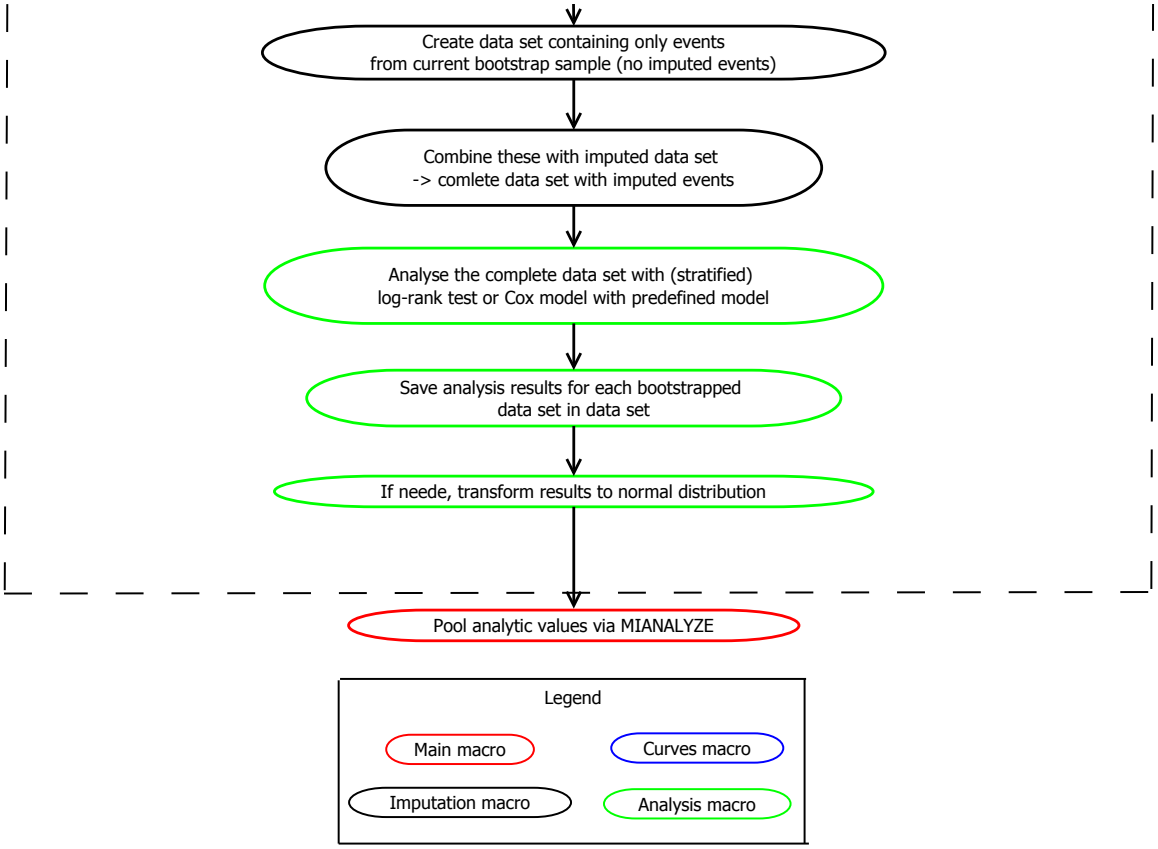
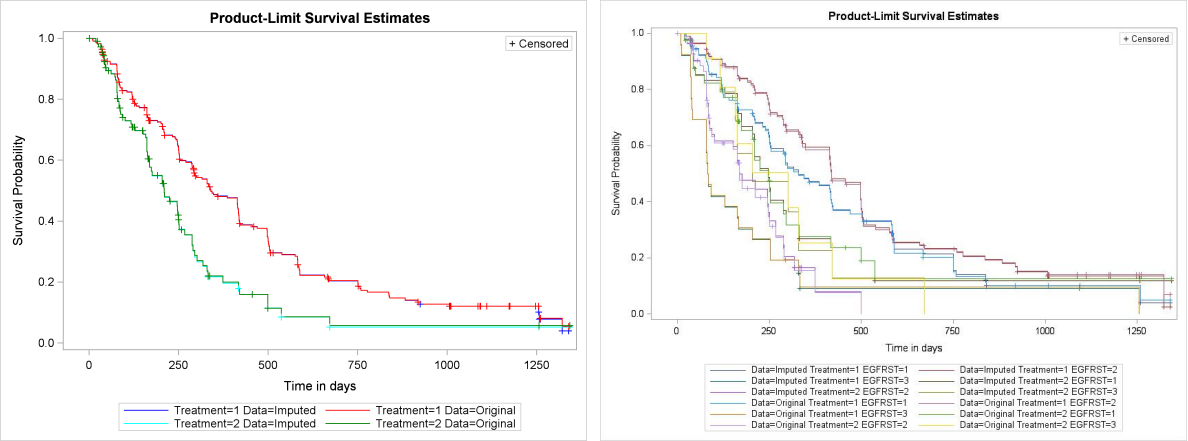


Figure 3.4: Flowchart of CAR macro

3.3 Results

The macro produces B imputed data sets from the original data set where B denotes the user specified number of bootstrap samples. Additionally it returns outputs of the *MIANALYZE* procedure.
The imputed data sets can be pooled to an overall data set with B times the population of the original dataset and compared to the original data set. This comparison is done to show that no bias is introduced by imputing the censorings. All imputations in this Thesis are done 100 times with $Seed = 1111$. The first approach is the comparison of the resulting Kaplan Meier curves.

Figure 3.5a shows the Kaplan Meier curves calculated on the basis of the imputed data set and the original data set respectively, both separated by treatment. The respective curves almost coincide. The jumps in the curves for the imputed data sets are located at the same time points as the ones in the original curves. This is caused by the fact that the imputation produces more events but not more event times, since the censorings are replaced by already existing event times. The survival times of censored observations that could not be imputed as events generally fall on the last 10-15 event times. The usually do not fall on earlier event times because the KMI only imputes as censoring if the generated random number has a smaller value than the smallest probability in the KM table corresponding to the last possible event time in the current KM table. The variation within the survival times that remain censored is due to the use of bootstrap samples which can have differing last event times.
The right figure 3.5b shows the curves after imputation separated by treatment and stratum. Those curves differ more clearly than in 3.5a. This can be explained on the basis of table 3.2 which is presenting the patient



(a) Imputation by treatment

(b) Imputation by treatment and stratum

Figure 3.5: Imputation under CAR

numbers per stratum.

	Treatment			Reference		
Stratum	1	2	3	1	2	3
Observations	91	112	27	47	57	11
Events	72	80	24	26	35	9
Censorings	19	32	3	21	22	2

Table 3.2: Number of patients by treatment and stratum

Especially from stratum 3 within treatment 2, it is observable that some stratum groups are very small and the difference between the curves of the original and the imputed data is a consequence of these small group sizes. However, the graphs are still very similar such that also the descriptive statistics prior and after the imputation shouldn't deviate much from each other. The Quartiles of the original and not stratified imputed data sets are displayed in table 3.3.

	Original		Imputed	
	Treatment	Reference	Treatment	Reference
25% Quartile	162	93	162	93
Median	340	210	340	210
75% Quartile	586	333	586	330

Table 3.3: Comparison of Quartiles of observation times under CAR

The numbers for both treatment groups coincide quite well which is another strong hint that the imputation preserves the distribution in the different groups.
The next step is to examine the analysis methods for TTE data which are implemented in the program.

	Hazard ratio				
	Value	Lower CI	Upper CI	Score	p-value
Original	0.5712	0.4283	0.7617	3.8128	0.0001
Imputed	0.5591	0.4140	0.7552	3.80	0.0002

	Log-rank test				
	Value	Lower CI	Upper CI	Score	p-value
Original	-23.115	-34.8580	-11.3719	-3.8589	0.0001
Imputed	-35.3103	-52.8746	-17.7460	-3.95	$<.0001$

Table 3.4: Comparison of analytical values under CAR

Table 3.4 presents the results of the hazard ratio and log-rank test without stratification for the original and the not stratified imputed data set. The scores for the original data are z-scores, whilst the ones for the imputed data sets are t-scores. If a hazard ratio is equal to 1 the hazard in the active treatment group is the same as that for the the reference group which means the active treatment shows no effect. A hazard ratio < 1 shows a reduced hazard in the active treatment group compared to the reference group.
The values of the hazard ratio are close to each other but those of the log-rank statistic differ clearly. The reason for this can be identified by looking at the formula of the log-rank test (chapter 2.1.5). As the test statistic takes into account the number of events and the imputed dataset has more events than the original one since most of the censored values have been imputed, this implies a difference in the denominator. The p-values of the log-rank test however can be compared after the pooling by the *MIANALYZE* procedure [35].
To ensure that the differences of the log-rank statistics are due to the increase of events and not caused by an error in the macro, an equivalence test was performed upon a set of simulated data.

3.4 Equivalence test

The aim of this test is to demonstrate that the results for the log-rank test are not biased by the conduct of an imputation process as described in this chapter.
A pair of simulated data sets were provided: the first does not contain any censorings; the second one is identical to the first data set but is censored at random with a censoring rate of approximately 20%. The data sets contain 100 sets of simulated trial data with 2000 patients (1000 per treatment group). Both the event times and the censoring times follow an exponential distribution but with different parameters. The parameters of the exponential distribution are chosen in order to end up with $\approx 20\%$ censorings in each of the 100 trial data sets. Those data sets were provided for usage in this thesis by an external source and not created by the author. Those two data sets offer the possibility to compare an imputed data set to its ”true” version.
The CAR macro was used to impute the censored data set. The results were then compared to those from the complete data set.
Since the use of the CAR macro will result in an imputed data set with $\approx 1\%$ censorings, the remaining censored values will be copied to the complete data set after every imputation. Hence, both data sets have the same number of events and therefore the log-rank statistics are comparable.
The log-rank statistic of the complete data set is calculated by means of the *LIFETEST* procedure after the remaining censorings are taken over from the imputed data set. The imputed data sets are analysed with the same procedure.
Finally, 100 log-rank statistics are obtained for the complete and the imputed data sets, respectively. To determine if there is a significant difference between both those statistics, a two one-sided t test (TOST) is performed via the *TTEST* procedure.

The mean of the LR statistics for the complete data sets is -356.85. Therefore, the margins are defined as $\theta_L = -3.5$ and $\theta_U = 3.5$ for the TOST in this case because this allows approximately 1% divergence for the mean of the LR statistic for the imputed data set relative to the means of the log-rank statistics of the complete data set.
The summary of the mean difference and the results of the TOST is displayed in tables 3.5 and 3.6.

Mean	95% CL		Std Dev	Minimum	Maximum
-1.0419	-2.9299	0.8461	10.3293	-20.3197	23.7095

Table 3.5: Summary of the mean difference between log-rank tests of imputed and complete data sets

The difference between the means of log-rank tests of the different data sets is -1.0419 which is very small in relation to the absolute values of the test statistics. The fact that the confidence interval includes zero reassures that the statistics are not different, but does not eliminate the possibility that there is simply insufficient data.

Mean	Lower Bound		90% CL			Upper Bound	Assessment
-1.0419	-3.5	$<$	-2.6218	0.5380	$<$	3.5	Equivalent

	H0	DF	t-Value	p-value
Upper	-3.5	99	2.58	0.0056
Lower	3.5	99	-4.77	$<$.0001
Overall				0.0056

Table 3.6: TOST Level 0.05 Equivalence Analysis

Table 3.6 shows the results of the TOST equivalence analysis for the Null hypothesis $H_0 : \Delta < \theta_L$ or $\Delta > \theta_U$ vs. the alternative hypothesis $H_A : \theta_L < \Delta < \theta_U$ where Δ denotes the difference in the LR test statistic and θ_L and θ_U are the predefined lower and upper equivalence margins.
The TOST procedure consists of two separate tests ($H_{01} : \Delta < \theta_L,\ H_{02} : \Delta > \theta_U$) and consequently, the test results comprises two p-values. The given results reveal that both null hypotheses can be rejected and the overall p-value of 0.0056 also signifies the equivalence of the LR test statistics at a significance level of 5%.
The theory for the TOST is taken from [26] and [32].

This equivalence test shows that the difference observed in the log-rank statistic can probably be explained by the increased number of events in the imputed data set. If a log-rank test is used the comparison has to be done with the p-values.
Table 3.7 reveals that the hazard ratios of the censored, the uncensored and the imputes data sets are very close to each other as well as the t-scores and the p-values. For the log-rank test, the censored data set which has less events has a clearly different log-rank statistic. However, the t-scores and p-values are close to each other and can be compared.

	Hazard ratio				
	Value	Lower CI	Upper CI	Score	p-value
Original censored	0.4265	0.3663	0.4965	11.01	<.0001
Original uncensored	0.4291	0.3724	0.4945	11.73	<.0001
Imputed	0.4264	0.3640	0.4996	10.61	<.0001

	Log-rank test				
	Value	Lower CI	Upper CI	Score	p-value
Original censored	-287.6332	-334.765	-240.501	-11.99	<.0001
Original uncensored	-356.846154	-408.093	-305.599	-13.68	<.0001
Imputed	-357.8880	-412.221	-303.555	-12.95	<.0001

Table 3.7: Comparison of analytical values for equivalence

This program also shows that the CAR multiple imputation program works correctly and produces results that can be considered as a complete data set with the same distribution as the censored data set.

Chapter 4

Informative censoring

The previous chapter described the structure of a Macro which imputes censored values under the assumption of CAR. By making some changes, this program can also handle informative censoring. Three different approaches to do this were realised in this thesis: tipping point analysis, reference based-imputation and pattern imputation.

4.1 Tipping point analysis

The Kaplan Meier curve estimates the probability of not having the event of interest up to a certain time point. In case CAR is assumed, a censoring only causes a reduction of the number of patients under risk at the time of drop out.
For the following sensitivity analyses, it is assumed that individuals in the treatment group who drop out of the study would actually do worse than the average of individuals in this group. This assumption is based on the thought that patients leave the study because they don't feel any benefit from the treatment and therefore either think the treatment is not working properly or that they may have been assigned to the reference group. That is, the censoring is not at random, but depends on the treatment effect.
The idea behind the tipping point analysis is to take each probability estimated for the Kaplan Meier curve to the power of a predefined factor δ. Since probabilities are always in the $[0, 1]$ interval, taking the probabilities to the power of δ leads to a reduction assuming delta is greater than 1 [14][35]. This δ can be interpreted as an increased hazard after the patients drop out of the study. For example, if $\delta = 2$ the hazard for an event at time t is twice as high compared to the patients who remain in the study [35].
The goal of the tipping point analysis is to find the smallest value of δ that worsens the Kaplan Meier curve of the treatment group such that the difference between the survival curves of the treatment and the reference is no longer significant. If the δ does not seem clinically reasonable, the results can be called robust to the CAR assumption [14].The implementation of this approach only demands slight modifications of the CAR macro. In advance of the actual imputation process, all survival probabilities in the KM tables of the active treatment groups need to be taken to the power of the predefined δ. All groups apart from the reference group will be adjusted with the same δ. This is especially important if the imputation is done separated by stratum or if there is more than one active treatment group. In spite of these changes, the basic CAR imputation can still be reproduced by setting δ equal to 1.

4.1.1 Implementation

The implementation of this δ in the program is done via the previously mentioned *adjustment* macro. This macro is able to realize all the changes that are necessary for the tipping point analysis as well as for the next approach. It is called directly after the Kaplan Meier curves have been estimated for the respective data set. The changes are always done only to the Kaplan Meier curves which are used to find the imputation time, and not to curves calculated for an analysis.
The former CAR program that now includes the *adjustment* Macro can be called via a parental Macro which increases the δ for every run stepwise until the p-value exceeds a pre specified significance level or a maximal

number of runs has been reached. The call statement of the tipping point Macro differs only slightly from the one for the *main* macro.

```
%TP(    Data=lib.os_strata,
        level=0.05,
        maxdelta=7,
        start=1,
        by=1,
        Separate="NO",
        Analyse="NO",
        Stratum=EGFRST,
        Treatment=RANDTRTC,
        censor=eptcen,
        btloop=100,
        seed=1111,
        bt="YES",
        test="HR");
```

The *level*= statement allows the user to specify a specific 2-sided significance level for the treatment effect which is by default 0.05. The *maxdelta*= argument offers the possibility to have a maximum δ at which the tipping point analysis stops. This is needed because it is not guaranteed that the difference between the treatment groups can fall below being significant even for an infinite δ. The *start* statement is used to specify an initial value for δ and the *by*= option allows the user to define the step width for the sequential increase of δ. The value for δ does not need to be an integer, it can have any decimal number. The recommendation is to run the macro first only with integers and after the tipping point is identified to be in the interval $[x, x+1]$ the macro can be run again starting with a starting value x, a maximum of $\delta = x+1$ and a step width of 0.1. Thus, the tipping point can be found correct to one decimal place. As before, all imputations were done 100 times. For the tipping point analysis each result for one δ is calculated using 100 data sets.

4.1.2 Results

In order to find a δ which makes the difference between treatment groups not significant, the macro is applied to a subgroup of the original study. The reason why the original data set could not be used entirely will be explained later on in this chapter. The subgroup consists of patients with EGFRST equals 1 or 2. This does not have any clinical reason it is only to be able to present the functioning of the tipping point analysis. The resulting study population consists of an overall of 207 patients. The active treatment group includes 139 patients, 55 of them censored. The reference group consists of 68 patients with 18 of them dropped out. This subgroup is only used to demonstrate the functioning of the tipping point analysis.
Before showing the results of the tipping point imutation the difference between the Kaplan-Meier curves for the KMI are displayed for several δ's.
Figure 4.1 displays the Kaplan Meier curves of the active treatment group taken to the power of $1, 3$ and 5. To give an example of how much the choice of δ affects the curves, a dotted line is drawn at 50%. For $\delta = 1$ the intersection of the KM curve with this line is at day 339 which can be interpreted as the probability of not having an event until this day is 50%. For $\delta = 3$ and $\delta = 5$ the 50% mark is reached at days 122 and 79, respectively. Furthermore, one can observe that the curves for $\delta = 3$ and $\delta = 5$ seem to be equal to 0 which is

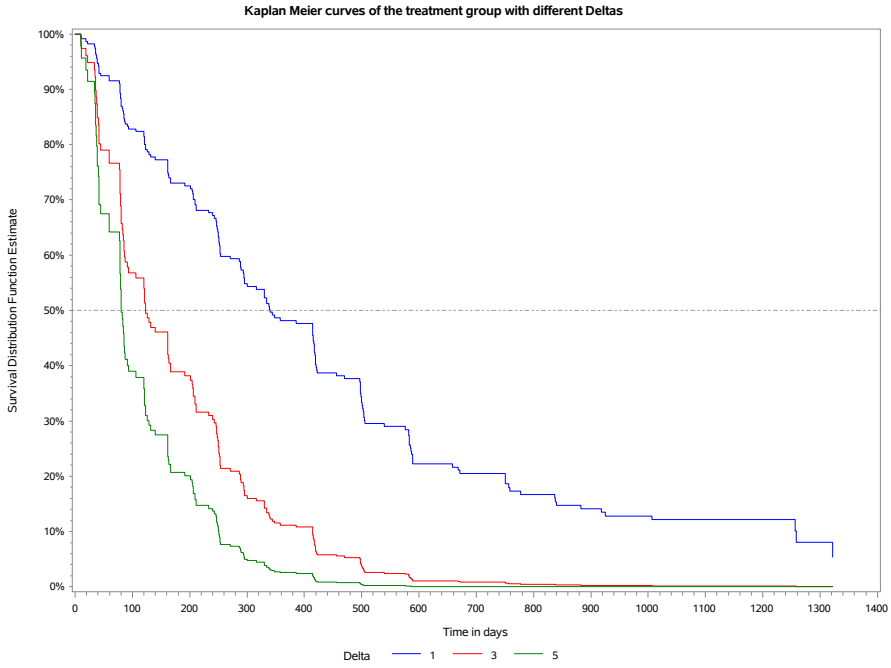


Figure 4.1: Kaplan Meier curves of the treatment group with different choices of δ

not possible from a mathematical point of view. In fact the probabilities for those curves are close to zero but not equal. If this were the case it were caused by the machine epsilon.
For the tipping point analysis the imputed observations are combined with the observed events which have event times according to the KM curve with $\delta = 1$.

The decrease of the difference with an increase of δ of the actual tipping point analysis can be seen in figure 4.2.
The reference curve is the one for $\delta = 1$ which is based on the the CAR imputed data set and not the original data set without the imputations. Since taking the Kaplan Meier KM table to the power of 1 does not change anything of it, the imputation with $\delta = 1$ is exactly the same as the CAR imputation explained in chapter 3.
The first thing to notice is that independent from the δ, the curves for treatment 2 (the reference group) are all identical. This was to be expected since the reference group is not affected by the tipping point analysis and the imputations were done with the same seed.
The changes only affect the treatment group which can also be seen in this figure. With increasing δ, the distance between the survival curves of the active treatment group and the reference group diminishes. However, the difference between the curves corresponding to $\delta = 1$ and $\delta = 3$ is much bigger than the difference between the curves for $\delta = 3$ and $\delta = 5$. This is caused by the fact that the curve which is used to generate the imputation times rapidly converges to zero as delta increases.

From the results for the HR and the p-value of the LR test, displayed in table 4.1, it can be deduced that the difference in the survival curves between the active treatment group and the reference group is no longer significant for a $\delta = 5$.

Generally, it can be observed that for an increasing δ the hazard ratio decreases and hence the p-value increases.

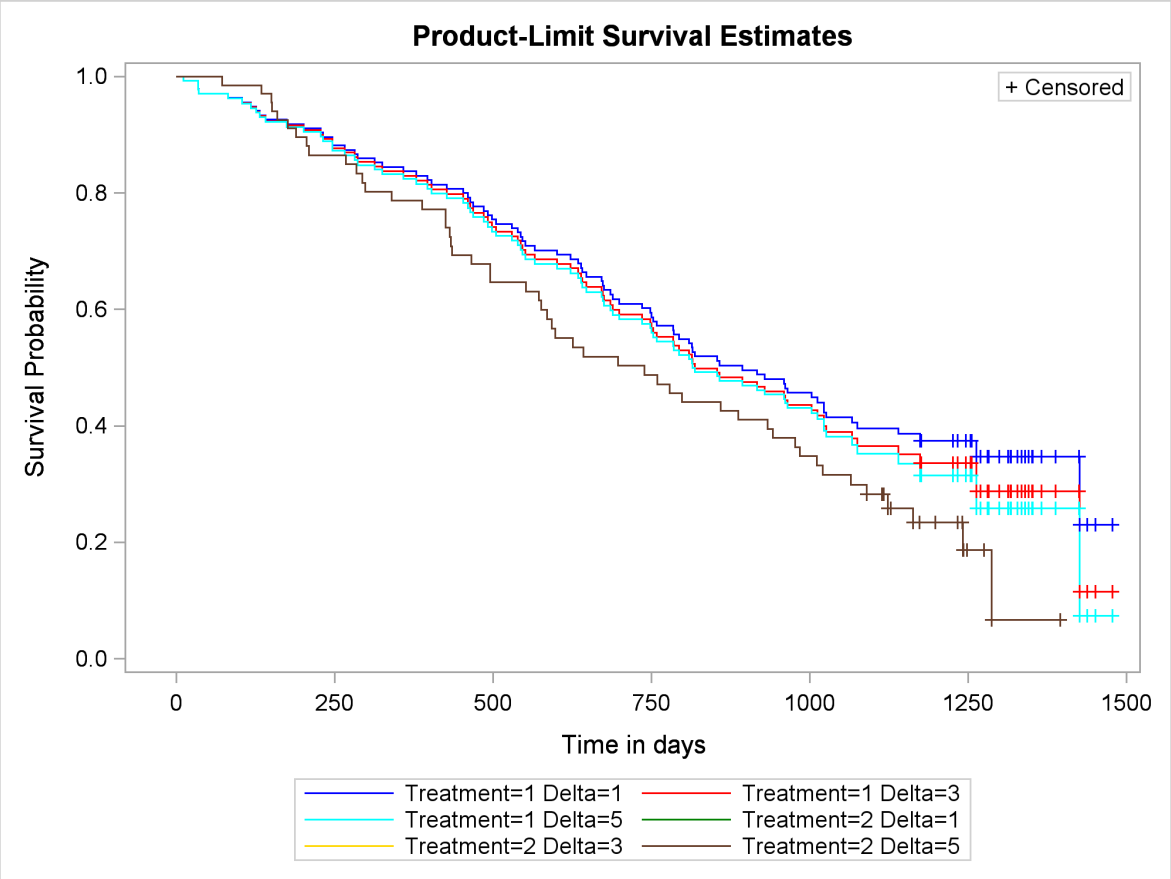


Figure 4.2: Kaplan Meier curves with $\delta \in \{1, 3, 5\}$

Delta	HR	Std Err	t value	p-value
1	0.6081	0.8252	2.59	0.0097
2	0.6466	0.8273	2.23	0.0216
3	0.6748	0.8294	2.05	0.0357
4	0.6945	0.8305	1.93	0.0497
5	0.7101	0.8315	1.86	0.0637

Table 4.1: Results from the tipping point analysis

The procedure stops when the p-value exceeds the significance level of $\alpha = 0.05$. However, this table only shows that the true tipping point lies in the range of 4 and 5. As stated before, the program can be executed a second time with *start*=4, *maxdelta*=5 and a step width of *by*=0.1. The results for this second run are displayed in table 4.2.

Delta	HR	Std Err	t value	p-value
4	0.6945	0.8305	1.93	0.0497
4.1	0.6957	0.8306	1.96	0.0507

Table 4.2: Results from the tipping point analysis to one decimal place

The p-value for $\delta = 4$ is already very close to 0.05 such that the true tipping point can be anticipated at a value close to that. The program identifies the tipping point for this data at approximately $\delta = 4.1$. It is approximately

because there are several steps where a random number generator is used (e.g. bootstrapping and drawing the random uniform [0,1] value). If the seed value is changed there will be a small variation in the results. The next step would be to evaluate if this value seems to be clinically plausible [14]. Since this is not a statistical issue and furthermore, it varies from case to case, this question will not be dealt with in this thesis.
The change of the estimated survival curve for the active treatment group with $\delta = 4.1$ from the Kaplan Meier curve for $\delta = 1$ is illustrated in figure 4.3.

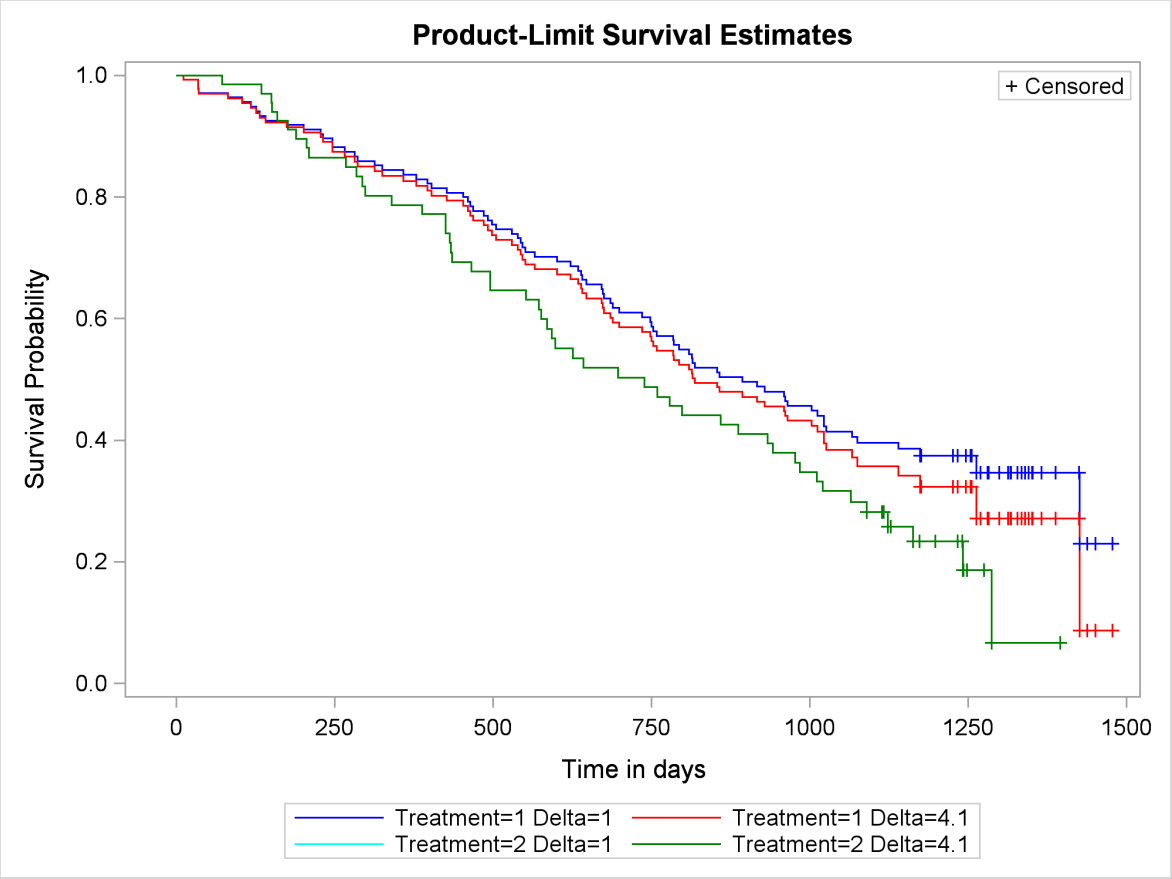


Figure 4.3: Kaplan Meier curves with $\delta \in \{1, 4.1\}$

The difference between curves for the active treatment groups (treatment=1) does not seem very big but suffices for the difference between the treatment and the reference group to turn no longer significant. The $\delta = 4.1$ can be interpreted as patients of the active treatment group who drop out of the study have a 4.1 times increased hazard compared to patients remaining in the study. This signifies a hazard ratio of $1/4.1 = 0.244$ for patients with events during the study compared to patients who were censored throughout the study.
An imputation separated by stratum and treatment is also possible to be investigated via the tipping point analysis. But since the data set used for the previous analyses consists of a stratum group of the data set introduced earlier there would be no difference to the already shown tables and graphs.

4.1.3 When no tipping point exists

As mentioned before, there is no guarantee for finding a δ which makes the difference between the treatment groups become no longer significant. An example can be given by using the complete data set introduced in

section 3.1. Figure 4.4 shows the Kaplan Meier curves for δ equal to 1 and 100.

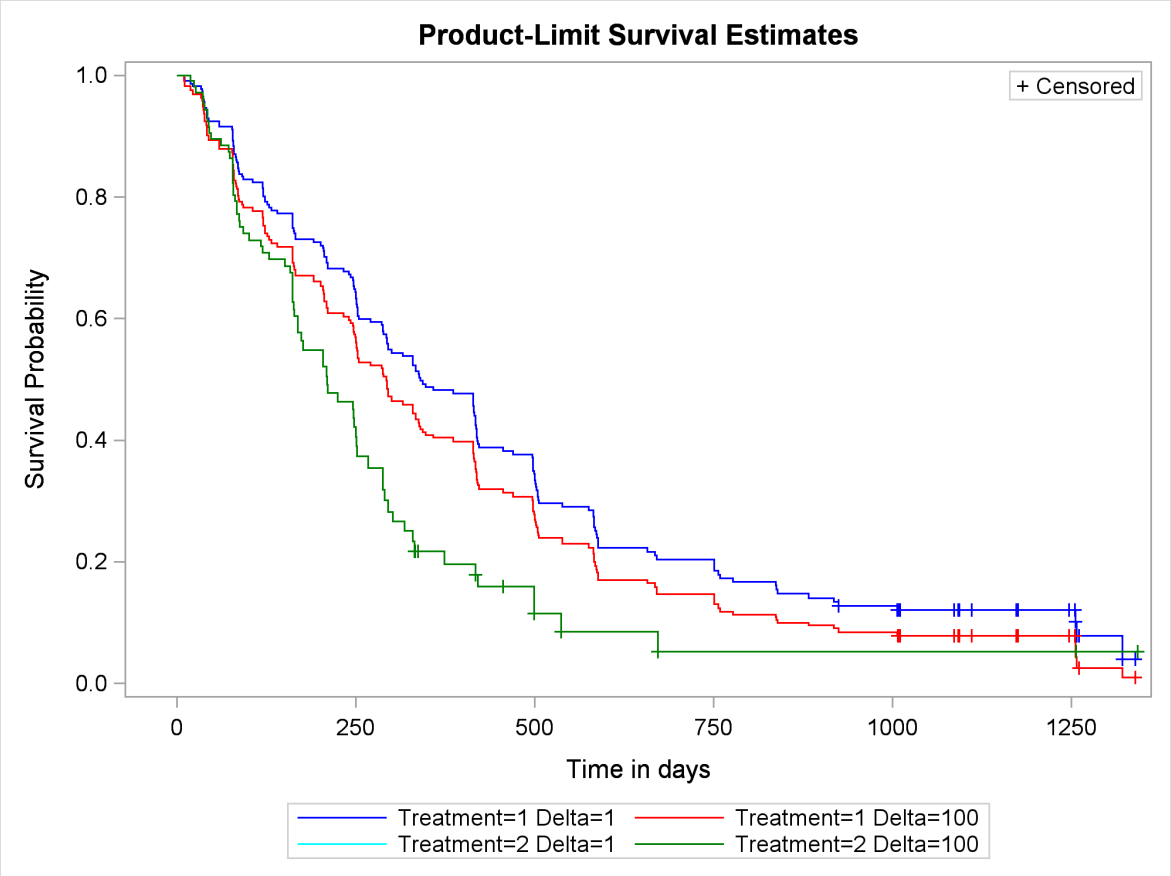


Figure 4.4: Kaplan Meier curves with $\delta \in \{1, 100\}$

The difference between the curves for $\delta = 1$ and $\delta = 100$ is very small compared to the increase of the δ. The reason for this behaviour is the convergence of the curves used to find the imputation times towards 0 for an increasing δ. If the δ could be set to infinity, the imputation time would be the time point for the next event (by treatment and/or stratum) after the observed censoring. Rothmann et al. and Zhao et al. call this the worst-comparison analysis [21] [35]. To get an impression of how this curve could roughly look like all censored values were imputed at the same day where the censoring happened, only the censoring variable is changed from 1 (censoring) to 0 (event).

These curves can not be produced by the program since the censored survival time is always imputed by an event time later than the censoring. Figure 4.5 shows how the Kaplan Meier curves would look like if all censored values would be imputed by an event at the day of the censoring. Again, this is not a realistic scenario but it can give an impression of how close the survival curves could get at the extreme. The hazard ratio can provide more detailed information about the difference between the treatments. The results are listed in table 4.3.

	HR	Std Error	Lower CI	Upper CI	t value	p-value
Treatment 1 vs. treatment 2	0.7657	0.8684	0.5807	1.0101	1.9057	0.0586

Table 4.3: Analytical values for setting censorings as events

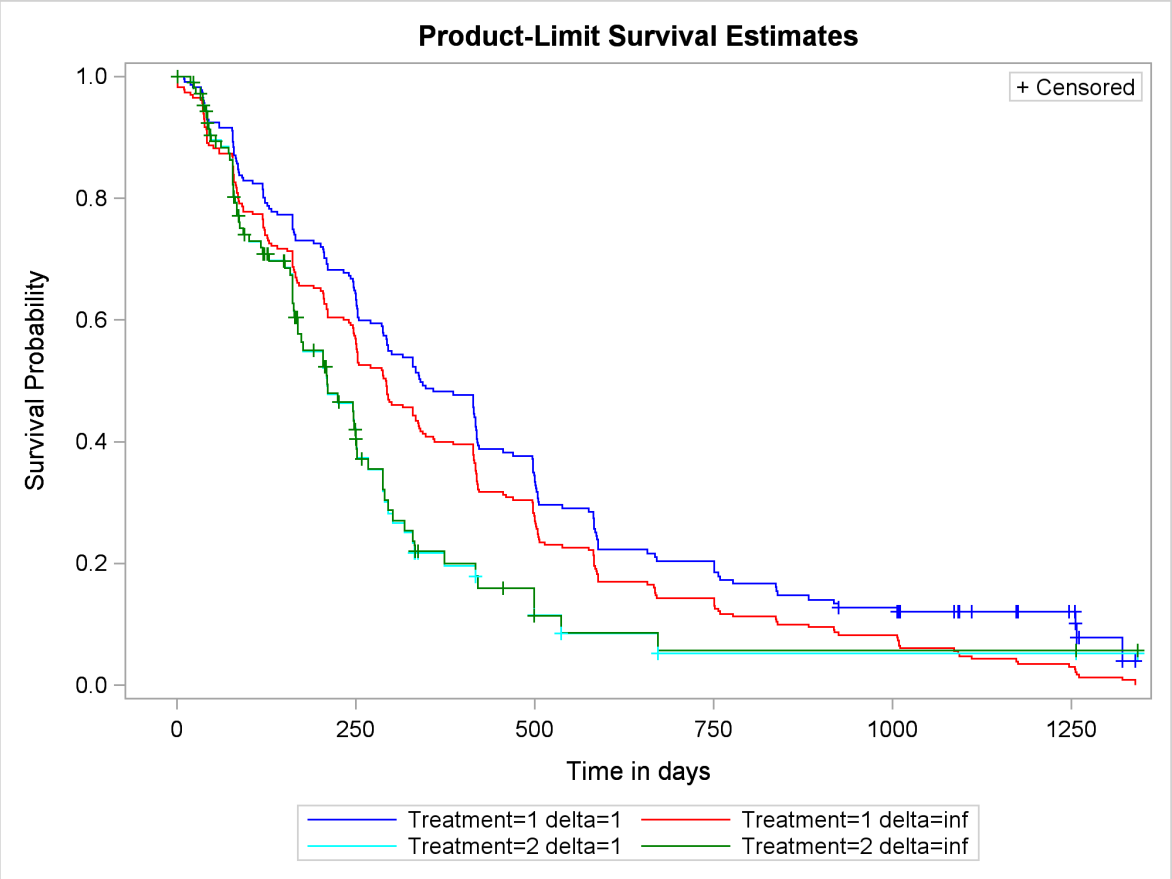


Figure 4.5: Kaplan Meier curves with censorings for the active arm imputed as events at day of censoring

The most important observation of this table is the lower confidence bound. The CI only just includes 1. The test if the point estimate for the HR is different from 1 has a p-value of 0.0586 so the difference between the two treatment groups is barely not significant with a significance level of 5%. Having in mind that this case cannot be handled by the program, it is highly doubtful that the tipping point analysis could overturn the conclusion of significance.

In conclusion, tipping point analysis can be used if the censored values are expected to have a true mean event time that is smaller than the actually observed mean event time in the same group. The δ is increased as long as the difference between the treatment groups is significant. If a δ results in a difference that is no longer significant, it is called the tipping point. If this tipping point is not clinically reasonable, the results may be judged robust against the CAR assumption [14].

It is not always possible to determine a tipping point since it might be the case that there is no δ such that the difference between the survival curves is no longer significant. This can be due to a small number of censorings or a strong treatment effect.

4.2 Reference-based imputation

Whereas the tipping point analysis presented in the previous section is a rather theoretical approach, the reference-based imputation introduced in this section is more based on practical considerations.

If patients drop out of a study and therefore are censored they will no longer receive the active treatment any more. For those observations, it is assumed that from this moment on they behave like patients that were as-

signed for the reference group, that is as if they are administered the reference treatment (this is because in oncology, the reference treatment is typically the existing standard of care). This assumption serves as a basis for the reference-based imputation. Hence, the values which are used to impute the censored observations are obtained from the estimated KM curve for the patients in the reference group. The idea for this assumption is taken from [20] [19].

4.2.1 Implementation

Only small changes in the introduced program are needed for the realization of the current approach. In the call statement of the *main(·)* macro only one additional option has to be added: the *control*= statement. If it is defined as *NO* the CAR imputation is done as before. Changing it to *YES* will activate the imputation process according to the above mentioned assumtions, implemented in the *adjustment* macro. As first step, the Kaplan Meier table for the reference group is duplicated and replaces the KM tables for all other the treatment groups. Only the *strat* variable is a combination of the treatment and stratum is maintained for each group. Figure 4.6 is a schematic representation for an example with 2 treatment groups only that illustrates which KM curves are used for the imputation of the censored observation on the two treatment groups.

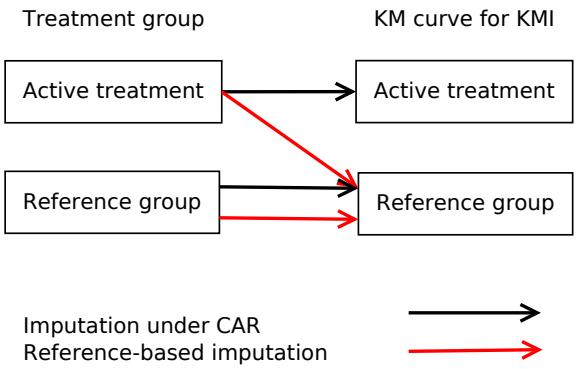


Figure 4.6: Schematic representation of the reference-based imputation

The schema shows that the censored observations of both, active treatment and reference group are imputed according to the KM curve which is estimated on the basis of the events in the reference group only.
If the imputation shall be done for each stratum separately, the KM tables for the treatment groups within one stratum are replaced by the KM tables of the reference group in this very stratum.

4.2.2 Results

For the comparison of the results obtained by the reference-based imputation vs. the results of an imputation under CAR, the final KM curves of both methods are plotted in a common plot (figure 4.7).
The difference between the KM curves for the reference group is negligible for the reason that the imputation process for this group is the same for both approaches. When comparing the curves for the active treatment group, one can observe that some of the jumps in the curve resulting from the reference-based imputation are not existing in the curve for the imputation under CAR for the active treatment group. This can be explained by the fact that the Kaplan Meier curve used to find the imputation times for the active treatment group is

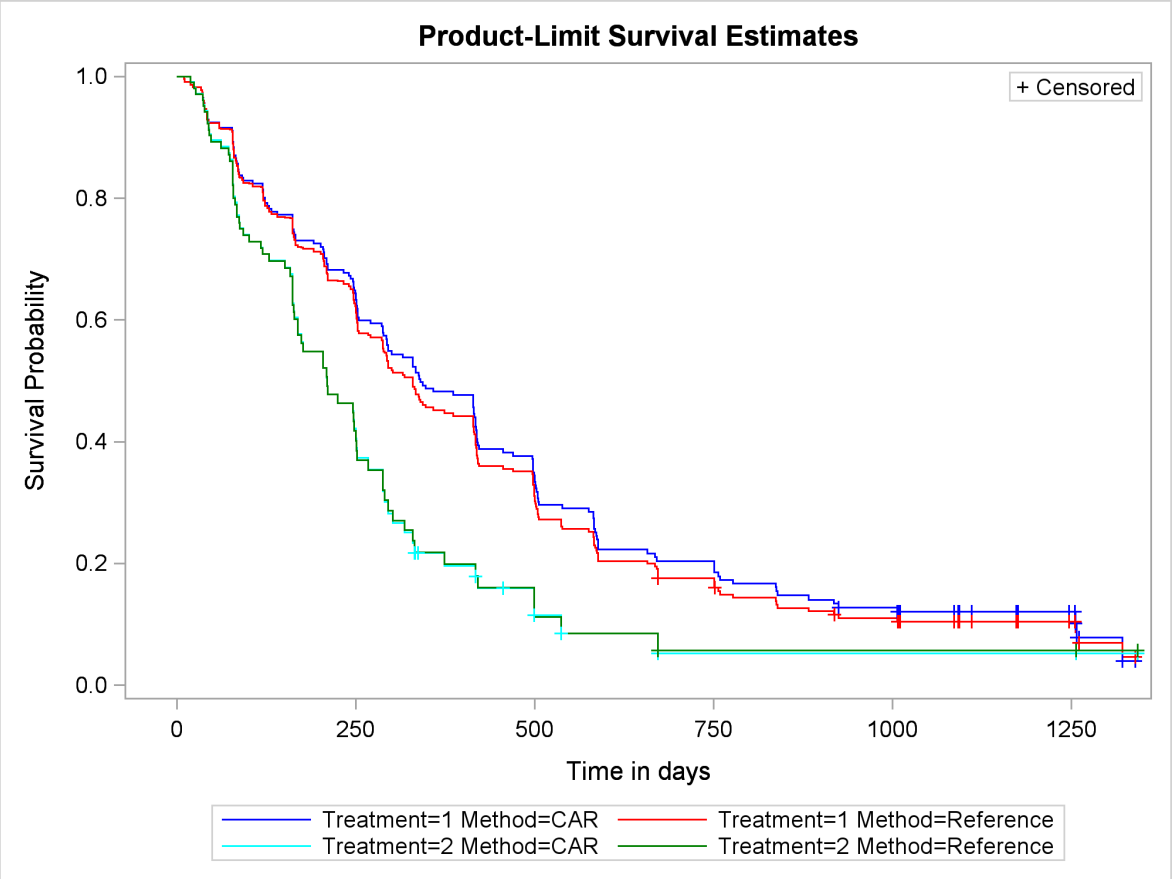


Figure 4.7: Comparison of the final KM curves for reference-based imputation and imputation under CAR

calculated from the reference group. This means a jump in the Kaplan Meier curve for the reference-based imputation is either an actual event observed in this group or a censoring which has been replaced with an event observed in the reference group.

As the difference between the curves produced for the tipping point analysis with $\delta = \infty$ and the ones for the imputation under CAR was much larger and still the difference between the survival curves of active treatment and reference group wasn't significant, it is not surprising that also the survival curves for the two treatment groups do not differ significantly after the reference-based imputation has been applied. The results may be seen in table 4.4.

Data	Hazard ratio	Lower CI	Upper CI	t value	p-value
CAR	0.5591	0.4140	0.7552	3.80	0.0001
Reference	0.5993	0.4510	0.7966	3.53	0.0004

Table 4.4: Comparison of analysis results for original data and reference-based imputed data

In comparison with the original data, the hazard ratio after the imputation is smaller and the p-value is bigger. However, the p-value still indicates a significant difference in the survival curves at a significance level of 5%. This shows that even if patients who drop out of the study are followed up until the end of the trial under the assumption that they receive the reference treatment and thus they are included in the final analysis, the difference between the treatment groups is still significant. Consequently, the primary analysis based on the one final data set can be considered robust against informative censoring.

A comparison between the reference-based approach and the tipping point approach can be drawn by finding a δ that produces approximately the same hazard ratio and p-value. Using $\delta = 1.57$ results in a HR of 0.6048 and a p-value of 0.0003, so reference-based imputation is approximately similar to a delta adjustment of 1.57.

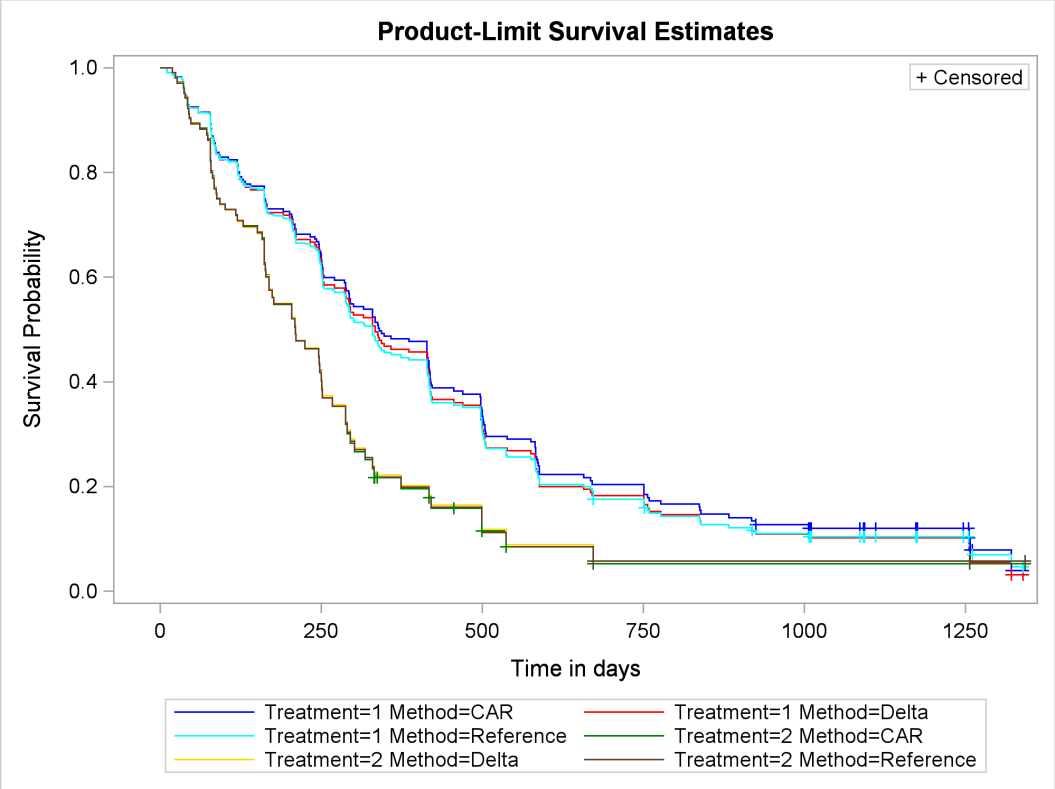


Figure 4.8: Comparison of final KM curves for reference-based imputation, delta adjustment with $\delta = 1.57$ and imputation under CAR

Figure 4.8 shows that the two curves for the active treatment groups obtained according to the reference-based imputation or the tipping point approach respectively are very close. The reason why the curves do not exactly coincide and would not be identical either if the hazard ratios would be equal is again that the reference-based imputation uses the Kaplan Meier curve of the reference group for the imputation of the censorings in the active treatment group.

Whereas the Kaplan Meier curves estimates the probability of not experiencing an event up to a certain time point. The reference-based imputation allows the illustration of an estimation of the reduction in the survival probability for a patient that drops out of the study for a reason related to the treatment effect compared to a patient that is continuing the study in the active treatment group 4.9.

One difference in the KM curve for the reference group compared to ones in the previous plots is that the curve ends at day 672. This is because the censorings are not taken into account when plotting the curve. Besides, the KM curve for the reference group is exactly the same as in all other graphs so presented.

The vertical line drawn at day 207 marks the point where the hypothetical patient drops out and consequently his survival probability is no longer estimated on the basis of the data for the active treatment group but on the data for the reference group. Up to that time the survival curve for the active treatment and the composite curve are identical and begin to differ only afterwards.

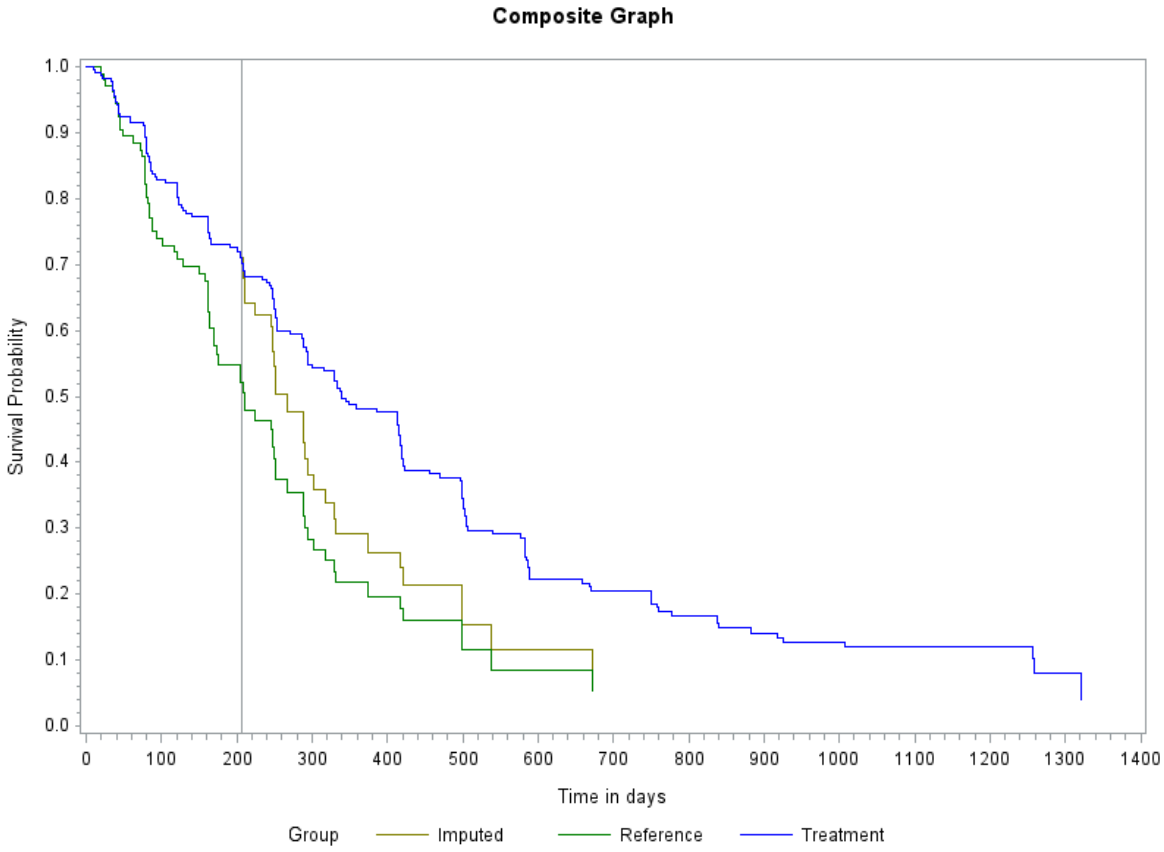


Figure 4.9: Composite graph compared to active treatment and reference under CAR

The curve referred to as ”composite” graph can be interpreted as the hypothetical Kaplan Meier curve of a patient who drops out of the study at day 207 and from then on receives the reference treatment. Its ’Hypothetical’ character results from the fact that the curve starts at day 0 when it is neither known whether the patient will be having an event or a censoring nor if it will be at that exact day 207. If it would be known for a patient that he is censored at that day and a new Kaplan Meier curve is to be calculated, the curve would start at 1 again since the patient surely did not experience an event until the day of the censoring.
At the time of the drop out the survival probability in the reference curve is lower than in the active treatment curve. To be able to construct this composite graph the KM table of the reference group is multiplied such that the survival probability at day 207 is equal to the survival probability in the active treatment group at that day. The next step is to combine the KM table of the active treatment group up to day 207 with the stretched KM table of the reference group containing only observations after this day.

4.3 Pattern imputation

4.3.1 The censorings

The first two approaches presented in this chapter are based on assumptions that concern all of the censored values in the treatment group. Realistically, patients drop out of a study for various reasons. Forming patterns of patients who leave the study for the same reason can give a more realistic idea of how the data might have looked without censorings.

The data set used in this thesis includes the following reasons for censorings:

1. No post-baseline imaging, alive and no progression during the trial
2. No post-baseline imaging, death or progression after second scheduled imaging
3. New anti-cancer therapy
4. Two or more consecutively missed images immediately prior to death
5. Alive and no progression at time of analysis

The first two cases already always occur for censorings at day 1 which means the patients were recruited and randomized but did not participate at any further examination. In the fourth case, it is known that the patient has died during the study but since two or more follow-up examinations are missing prior to the documentation of the death, the exact time of death or progression is not known. The last reason for a censoring is that the patient did not have an event or progression until the end of the study.

Reason	Overall		Treatment		Reference	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
1	6	6.06	2	3.70	4	8.89
2	6	6.06	2	3.70	4	8.89
3	61	61.62	28	51.85	33	73.33
4	2	2.02	2	3.70	0	0
5	24	24.24	20	37.04	4	8.89

Table 4.5: Frequencies of censorings by cause overall and separated by treatment

Table 4.5 presents the distribution of the censorings by their causes. With 61 out of 99 censorings over both treatment groups, a new anti-cancer therapy is the most frequent reason for drop out. The second biggest group consists of patients that did not have an event in the ongoing study.
Before giving examples of forming patterns for those censorings, the imputation process for the different patterns is explained and its implementation on the basis of the previously introduced programs is described.

4.3.2 The instruction data set

The *curves* and the *imputation* macros remain unchanged.
For the implementation of the previous approaches it was sufficient to specify one additional parameter (δ for the tipping point approach and *control=YES* for the reference-based imputation) to be able to do the imputation. For the imputation by patterns, one can specify a different δ for each pattern and every treatment group within this pattern. Alternatively, it is possible to select the Kaplan Meier curve for an arbitrary treatment group in the study to impute a specific pattern. And the program is able to use any Kaplan Meier curve and additionally apply a δ to this curve specifically for each pattern.
To make this possible without changing the call of the macro too much every time a different data set or different pattern shall be used, an instruction data set is created. For each pattern this data set contains the value for δ and the code for the treatment group which is used to calculate the Kaplan Meier curve for the imputation in a separate row. An example of how this data set could look like in the case of only two patterns is given in table 4.6. In general, the number of rows in the instruction data set equals the number of different patterns for the imputation.

	Treatment 1		Treatment 2	
Pattern	*Delta1*	*Switch1*	*Delta2*	*Switch2*
1	2	1	1	2
2	1	2	1	2

Table 4.6: Example for the instruction data set

The first row is just added for explanation and does not exist in the actual data set. The columns in the second row have to be named exactly as they are in the table since they are used by the program to identify the columns. In case there are more than two treatments, the numeration of the names is continued (Delta3, Switch3, Delta4, Switch4, ...). The "pattern" column also needs to be included in the data set so the program knows which line to use.
Treatment 2 is imputed similarly in both patterns: there is no delta adjustment added since $Delta2$ is equal to 1 in both cases and both patterns specify the KM curve of treatment 2 to be used for the imputation. This means the second treatment is imputed under the CAR assumption for both patterns as they are imputed using their own KM curve and $\delta = 1$.
The imputation of treatment 1 differs between the two patterns. In pattern 1 the KM curve of treatment 1 is used, so no switching of treatment groups is planned. But $Delta1 = 2$ specifies a delta adjustment for treatment 1 with $\delta = 2$ in pattern 1.
The second pattern includes no delta adjustment for treatment 1 ($Delta1 = 1$) but the imputation is planned to be done with the KM curve of treatment 2 ($Switch1 = 2$).
In summary, if treatment 2 is chosen as the reference group, pattern 1 is imputed by means of a delta adjustment with $\delta = 2$ whereas for pattern 2, a reference-based imputation is applied.

4.3.3 Implementation

The actual implementation of the pattern imputation is done by a modified version of the *main* macro, called *newmain*. The changes to the *main* macro are small but very important for the further course of the pattern imputation approach.
Apart from the original data, the *newmain* macro additionally imports the instruction data set and saves the number of treatment groups in the variable *nvar*.To regulate the order for the imputation process, a consequtive number is assigned to every stratum within each treatment group and pattern by means of the variable *"pattern"*. Table 4.7 takes up the example for the strat variable given in table 3.1 and illustrates the numbering for two patterns, treatment groups and strata.

	Pattern 1		Pattern 2	
	Treatment 1	Treatment 2	Treatment 1	Treatment 2
Stratum 1	1	3	5	6
Stratum 2	2	4	7	8

Table 4.7: Example of numbering contained in the *pattern* variable

Numbering the groups in this way enables the program to impute the patterns one after the other.
Another purpose of the newly created pattern variable is to specify the division of all censorings extracted from the original data set into the separate censoring data sets. In the *main* macro, this separation has been conducted according to the *strat* variable.
The macro for the adjustments for the pattern imputation approach is called *pmmadjustment*. It is able to make use of the instruction data set and the information about the number of treatments saved earlier in the variable

nvar. The *pmmadjustment* macro creates the data set *kmcurves* that contains all KM tables that are used for the imputation. Since every pattern can define the KM curve which has to be used for all treatments, the *kmcurves* data set contains as many KM curves as the number of treatment groups multiplied with the number of patterns, if no stratified analysis is done. In detail, for every treatment group in every pattern the *pmmadjustment* macro defines the KM curve of the treatment defined in the *switch* column as the KM curve for imputation for this specific pattern. After the KM table defined in the instruction data set is inserted as KM table for the *pattern* group it can be calculated to the power of the δ defined for this group in the instruction data set.

The program is able to impute the pattern appropriately as long as the instruction data set has a proper structure. The challenge for the user is to find patterns that are clinically reasonable. In the following three examples of patterns are given.

Pattern example 1

Table 4.5 reveals that the most frequent reason for drop outs is a new anti-cancer therapy. It may be assumed that patients from the reference group (treatment 2) receive a "better" treatment (that it is similar to the trial active treatment) and the patients from the active treatment group (treatment 1) a "worse" treatment than they did up to the drop out. Based on this assumption the instruction data set can be constructed as displayed in table 4.8.

	Treatment 1		Treatment 2	
Pattern	Delta 1	Switch 1	Delta 2	Switch 2
1	1	2	1	1
2	1	1	1	2

Table 4.8: Instruction data set for pattern example 1

Pattern 1 includes all censorings that drop out due to a new anti-cancer therapy. In this pattern, censorings in treatment group 1 are imputed using the KM curve from treatment 2 and vice versa. The second pattern includes all other censorings which are treated as censored at random by the imputation program.
Figure 4.10 shows the KM curves of this first example using a pattern imputation compared to the usual CAR imputed KM curves. The first thing to mention is that the curves of the reference group are different. In the previous approaches, the assumptions were focused on the active treatment group. The pattern imputation allows assumptions for the reference group as well. In this case it is assumed that patients in pattern 1 who drop out of the reference group receive the active treatment starting at the point of drop out. This change in treatment is the reason for the pattern imputed curve being above the curve of the CAR imputed reference group. In contrast to this, the curve of the pattern imputed active treatment group is below the curve of the CAR imputed active treatment group.
The log hazard ratio and other analytic parameters for the pattern imputed data set are displayed in table 4.9.

HR	Standard error	95% CL		t value	p-value
0.7175	0.8700	0.5460	0.9427	2.39	0.0172

Table 4.9: Pooled results from pattern imputation for pattern example 1

Since the imputed event times are better than the average in the reference group and worse than the average in the treatment group, the HR is smaller than in the original analysis of the data. The p-value of 0.0172 does indicate a significant difference between the treatment groups at a 5% significance level.

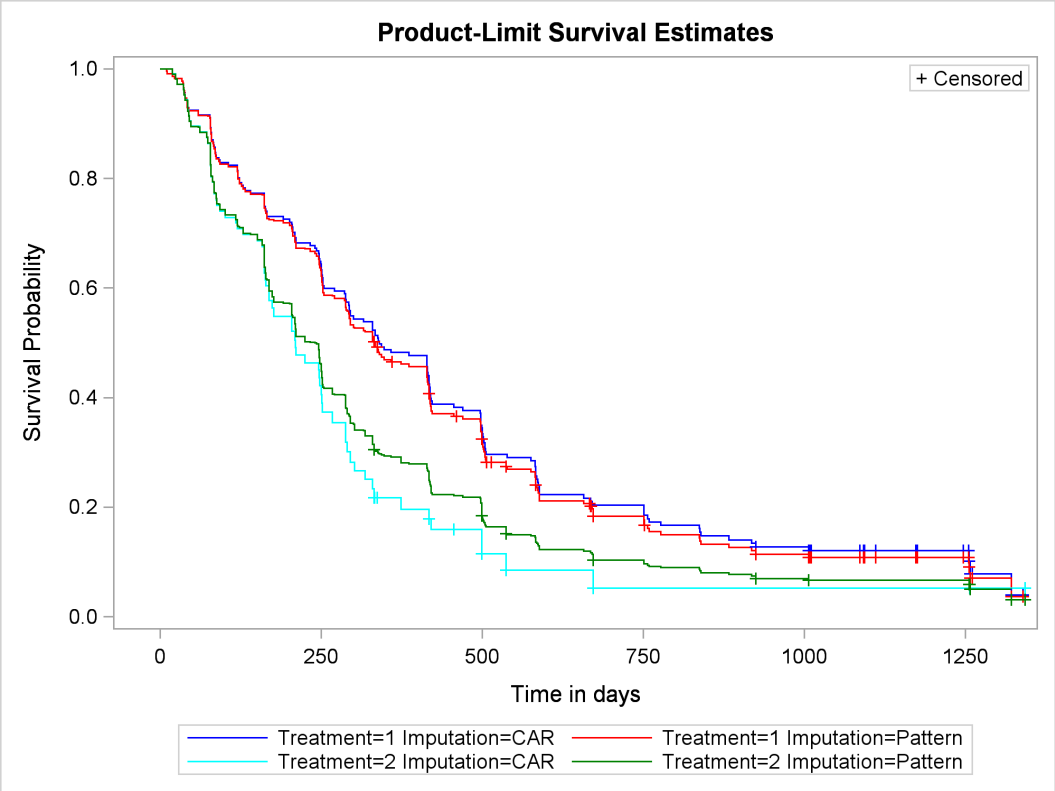


Figure 4.10: KM curves for pattern example 1 comparing pattern imputation with CAR imputed KM curves

Pattern example 2

The reason ”new anti-cancer therapy” is used again for another assumption to create patterns. This time, patients in the reference group will be assumed to be censored at random. The patients of the active treatment group which left the study for a new anti-cancer therapy are assumed to have the same risk as the patients in the reference group. Patients who were censored for other reasons are allocated to the second pattern and are imputed according to the CAR assumption. The instruction data set for this example can be found in table 4.10.

Pattern	Delta 1	Switch 1	Delta 2	Switch 2
1	1	2	1	2
2	1	1	1	2

Table 4.10: Instruction data set for pattern example 2

The difference from example 1 is that the censorings of the reference group in pattern 1 are imputed using the KM curve of the reference group and not the one from the active group.
Figure 4.11 shows the Kaplan Meier curves of the pattern imputed data and the CAR imputed data.

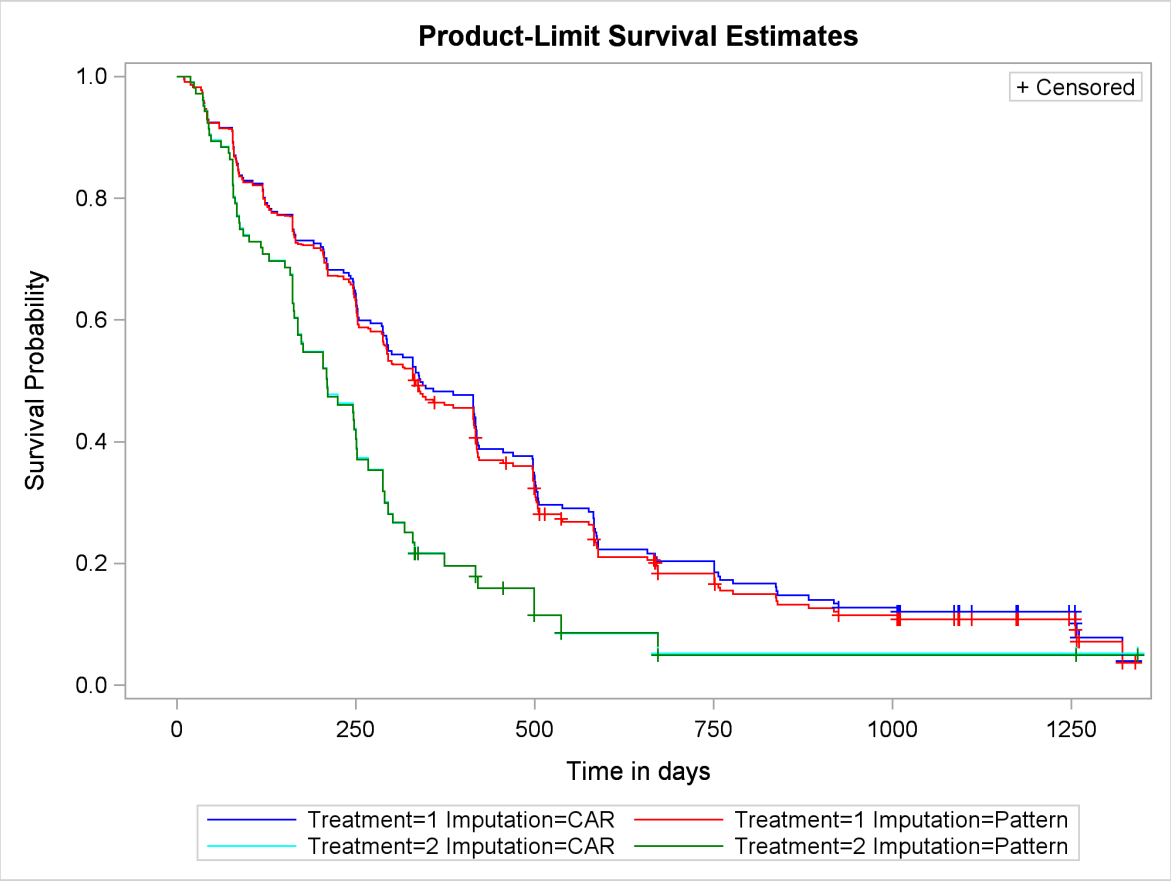


Figure 4.11: KM curves of example 2 of pattern imputation compared with CAR imputed KM curves

The small difference that can be observed between the curves of the reference group can be explained by the the fact that the pattern imputation imputes two groups of censorings separately from each other which leads to a higher variation in the imputation times as the KM curves for the imputation are based on a smaller number of observations.
The results from the *mianalyze* procedure are shown in table 4.11.

HR	Standard error	95% CL		t value	p-value
0.5842	0.8668	0.4413	0.7734	3.76	0.0002

Table 4.11: Pooled results from pattern imputation for pattern example 2

Comparing the log HR with the results of the analysis for the reference-based imputation (log HR of 0.5043, compare table 4.7) shows that the difference between the two treatment groups is bigger if only the censorings in pattern 1, are imputed using the reference-based method and the censorings in pattern 2 are imputed under the CAR assumption.

Pattern example 3

The third example is the most discriminating and demonstrates the flexibility of the pattern imputation. All causes for the drop outs are considered for setting up the patterns.
As before, the first pattern is formed by the drop outs due the administration of a new anti-cancer therapy. For this third example, the new therapy is assumed to have at best an equal effect equal to the reference treatment. This means that the censorings of all patients in this pattern, are imputed according to the KM curve of the reference group which is additionally worsened by choosing a small δ. The δ is applied because it is assumed that the disease of patients who drop out of the study to get another therapy is worse compared to the patients who stay in the study.
The next smallest group compromises the censorings due to reasons 1 and 5. In both cases, the patients are known to be alive at the end of the trial and were not diagnosed with a progression during the trial. Because of that, censoring at random is proposed: no adjustments to the KM curves are made. In the following description, it is referred to as pattern 2.
Pattern 3 is formed by patients who are censored because of death or progression after second scheduled imaging (compare reason 2 in section 4.3.1). This means that the patients are recruited for the study but no post-baseline imaging was carried out for them. The only available information is that death or progression has occured after the second scheduled imaging. The underlying assumption for the imputation is that those patients have a highly increased risk for an event compared to the treatment and stratum group they belong to. Therefore, a high delta (5) is applied to the KM curve of this pattern.
The last pattern consists of only two patients: those who missed two or more consecutive images immediately prior to death. Creating a separate pattern for two censorings may not influence the results too much but there is no minimum number of censorings that a pattern needs to contain. Since it is known that the patients in this pattern died during the study, again a high delta is applied to the KM curve of this pattern.

All these explanations lead to the instruction data set displayed in table 4.12.

Pattern	Delta 1	Switch 1	Delta 2	Switch 2
1	2	2	1	2
2	1	1	1	2
3	5	1	5	2
4	7	1	5	2

Table 4.12: Instruction data set for pattern example 3

Imputing the censorings according to these patterns results in the Kaplan Meier curves as shown in figure 4.12. Both curves of the pattern imputation are below the respective curves of the CAR imputation. The difference between the pattern imputation and the CAR imputation is larger for the active treatment group. Furthermore, for most of the patterns it is assumed that the patients who drop out of the study would do worse than the average of the uncensored patients. This is especially true for the patients in the activee treatment group. Therefore the influence of the imputed censorings on the KM curve of the active treatment group is stronger. Another reason is that for pattern 1, the censorings in the active treatment group are imputed using the KM curve of the reference group and additionally a δ of 2 is applied.
The results of the final analysis for the pattern imputed data set are displayed in table 4.13.
The results show a clearly significant difference in the survival curves at a 5% significance level. Specifying the patterns this way is complicated, requires time and clinical experience to figure out optimal specifications for each pattern. The advantage of this patterning is that the outcome of the imputation is potential realistic.

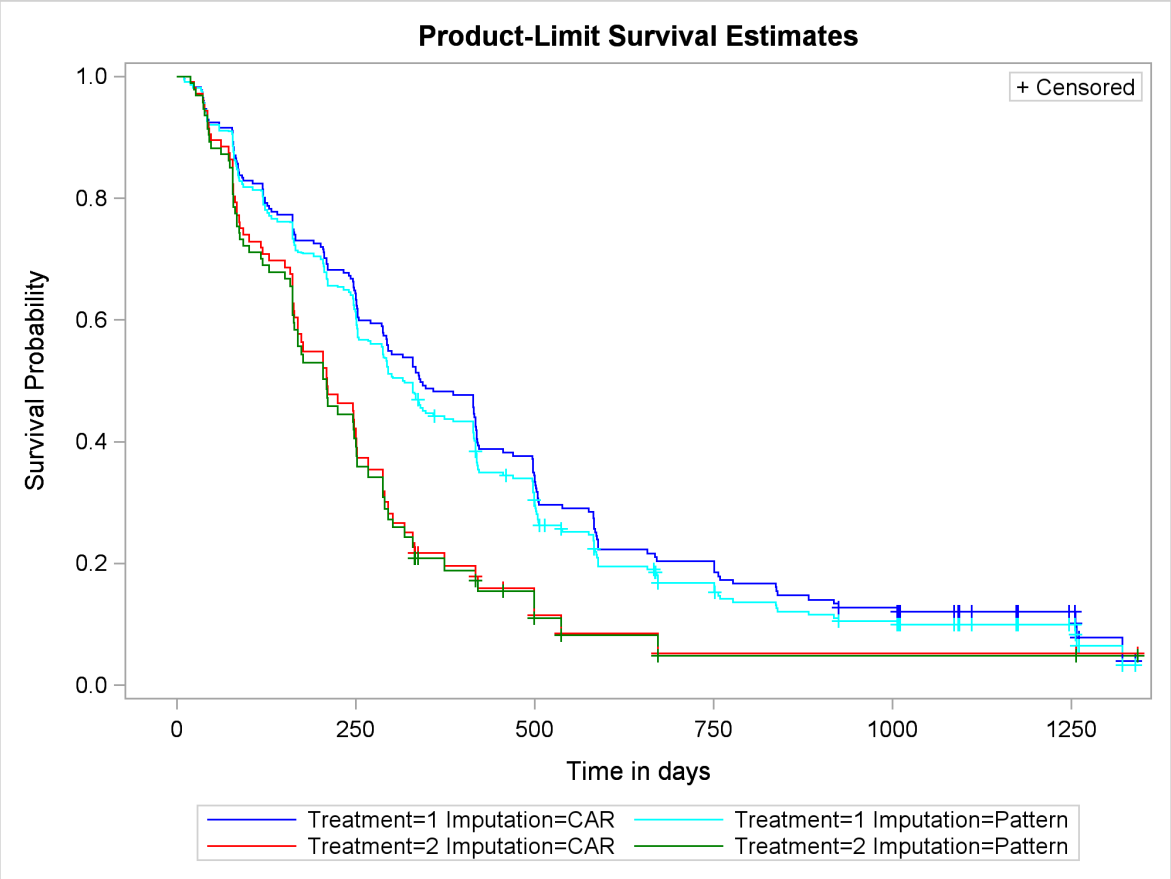


Figure 4.12: KM curves of example 3 of pattern imputation compared with CAR imputed KM curves

HR	Standard error	95% CL		t value	p-value
0.6007	0.8669	0.4538	0.7951	3.57	0.0004

Table 4.13: Pooled results from pattern imputation for pattern example 3

In conclusion, pattern imputation allows the application of imputation methods based on different assumptions to predefined patterns. The program written in the course of this thesis is not only able to apply specific δ's to different treatment groups within a pattern, it can also select an arbitrary Kaplan Meier curve that can be calculated on the basis of the provided data set for the imputation of a certain set of censorings.
With realistic assumptions based on clinical knowledge and reasonable patterns, the program gives an impression of how the data and the results might look like if no censorings had happened during the study.

Chapter 5

Discussion

This thesis takes up the idea of multiple imputation for time-to-event data under the censoring at random assumption and uses it as a basis for the development of sensitivity analyses that can be applied if informative censoring is assumed.
Multiple imputation procedures for time-to-event data are not very popular. The reason for that is that most analysis methods require the assumption of censoring at random. However, if multiple imputations are done under this assumption, the analysis of the imputed data will produce the same results as the analysis of the original data [30]. Hence, the imputation process doesn't add too much benefit but affords one additional working step and increases the standard error. Nevertheless, a program that implements multiple imputation under the assumption of censoring at random serves as a starting point for the main part of this thesis which is the implementation of 3 different approaches how to use multiple imputations for a sensitivity analyses to informative censorings in the statistical analysis software SAS. Independent from the respective censoring assumption, all implemented programs use the Kaplan Meier imputation introduced by Taylor et. al. [30] to generate the imputation times.
Generally, all programs enable the user to realize the implementation separated by treatment and possible stratification variables or by the treatment only. If using the implementation by stratification factors one has to be careful with the size of the stratification groups. If a group has not enough events the Kaplan Meier curve for the KMI of this group is poorly defined. This can lead to imprecise outcomes of the imputation and therefore imprecise results. Nonetheless, if there is a satisfying number of events in every group the stratified imputation uses more information about the data and the results are more precise.
After the implementation of the imputation under CAR has been accomplished, the code has been validated by comparing the produced results with the results of the original data. As expected, the estimated hazard ratios showed satisfying similarity. The small difference between the results of the original and the imputed data are assumed to be caused by the number of events. This assumption is based on the observations made for the equivalence test. For this test data sets with 2000 patients each were used and the differences between the hazard ratios of the original and imputed data were much smaller than in the example data set for this thesis which consists of 345 patients overall. If there would be an infinite number of events the Kaplan Meier curve would give a closer estimation of the Survivor function, assuming there is no bias, which is directly linked to the distribution of the events and therefore the imputation of the censored observations would be done according to exactly the same distribution like the censorings. This is no failing of the bootstrap since the bootstrap used in this thesis sampled the existing observations from the original data set and did not create new event or censoring times. Increasing the number of imputations will lead only to slightly closer results from a certain number of imputations on.
One reason for different results could be non proportional hazards as suggested by Zhao et al. [35]. In a study with a high censoring rate and a bigger treatment effect early on in the study the estimation of the hazard ratio may tend to be mainly influenced by events from the earlier part of the study. The Kaplan Meier imputation method puts more weight on the later part of the study by imputing potential event times. This can lead to hazard ratios with smaller effect for the active treatment group (closer to 1) after the imputation. This means that the Kaplan Meier imputation under the CAR assumption can be useful for the evaluation of implications of non proportional hazards during a study [35]. Since the difference between hazard ratios of the imputed and the original data set are relatively small there is no evidence of big violation for the proportional hazards as-

sumption. However, the results do not confirm the assumption, there is still the possibility of non proportional hazards in this data that could not be detected by the CAR imputation method.

Contrary to this, the values for the log-rank statistic diverge from each other. The reason was found in the fact that the test statistic takes into account the number of events. Since the imputation results in a higher amount of events, this causes the difference in the log-rank statistics. It may be possible to adjust the statistic by applying a factor related to the number of events before and after imputation to it. This is a possible topic for further investigations as it was not a part of this thesis finding such a factor.
However, an alternative is to compare the p-values of the log-rank test before and after the imputation as it is done in several publications [31][14][30]. The difference between the p-values can be explained with the same reasons for the differences between the hazard ratios. Additionally, this explanation has been confirmed numerically by conducting an equivalence test for a simulated data set with no censorings and the same data set for which some observations have been censored artificially and have been imputed afterwards by means of the written program. The boundaries of [-3.5,3.5] imply an allowed divergence of 1% relatively to the mean of the log-rank statistics of the non-censored data set. Since the TOST was made with 100 data sets the deviation in the respective groups is quite high which leads to a standard error of 0.952 for the mean difference between the log-rank tests of imputed and complete data sets. If the TOST would be done using more data sets to get a smaller standard error it may be possible to specify smaller boundaries to get more sensitive results. For this case boundaries with $\pm 1\%$ seem appropriate.
The test was able to show that the difference in the log-rank statistic before and after the imputation is mostly caused by the increased number of events after the imputation.
In general, the CAR imputation works and can be used to impute time-to-event data sets. Since this thesis was limited in time there are options that could not be included but are worth to be mentioned. The first point is concerning the analysis of the imputed data sets which does not offer the possibility to add covariates for the Cox proportional hazards model. Variables other than the treatment group can only be included as stratification factors and so, no parameter estimates can be calculated, so the influence of the variables cannot be measured.
For the imputation itself, the Kaplan Meier imputation is the only imputation method. Other imputation methods could be included in the program as well, for example the risk set imputation which has been introduced in section 2.3.2.
Another reason for differences in the results for both the hazard ratio or the log-rank test could be an insufficient number of imputations. All results in this thesis were produced using 100 imputed data sets. If no bootstrap is used this number should give stable results but a bootstrap brings a higher variance with it. A higher number of imputations should result in more stable point estimates and reduced variance. This could not be done in the thesis due to the long run time for higher imputation numbers than the used 100. Nonetheless, the use of the bootstrap makes sense since it gives a better estimate for the variance.
Since the Kaplan Meier imputation method is used one could think about implementing an option to "specify" an external Kaplan Meier curve to be used in the imputation process. This can be useful for example if patients drop out because they receive another therapy. In this case the survival times could be imputed according to the KM curve of this new treatment which can be taken from the respective trial. It could also be used if there is a subgroup of censored patients with characteristics that have been examined in another trial.
A problem with the Kaplan Meier imputation is that for the last censorings that occurred in the study only few events remain to calculate the corresponding Kaplan Meier curves or no events remain at all if the censorings occur after the last event. This can lead to imprecise results. Zhao et. al suggest an exponential model for the conditional survivor function for the last few events. For a censoring time c_k after the last event this would be $\hat{S}(c_k) = \hat{S}(t_M) * exp(-h * (c_k - t_M))$, where h denotes the hazard assumed for the censored observations

[35]. This could not be included in this thesis due to the restricted time but could be an interesting point for possible extensions of the implemented program.

The basic idea for the imputation of informative censoring is to modify the KM curves used for the KMI according to the assumptions for the informative censoring. Zhao et. al proposed to add the power of a $\delta > 1$ to the KM curve and thus to simulate a deterioration of the survival times of the drop outs [35]. This delta adjustment can be used as sensitivity analysis for informative censoring. O'Kelly used the delta adjustment to perform a tipping point analysis [14], which is the first approach realized in this thesis.
The delta adjustment method is useful if the difference between the treatment effects is small and informative censoring is assumed for the active treatment group. Since this method "searches" for the δ which worsens the active treatment group so much that it is no longer significant it would make sense to pre specify a δ with a medical expert as threshold for the sensitivity analysis. This is because the term "clinically reasonable" can be subjective depending on what needs to be shown.
Zhao et al. choose $\delta = 2.5$ as upper bound which corresponds to a hazard of 0.4 comparing treatment vs. placebo as represents a reasonable effect size for a clearly effective treatment for their example [35].
The main target of a tipping point analysis is to find a δ that makes the treatment effect no longer significant. This can not always be achieved as shown in section 4.1.3. But even this scenario gives information, which is that the treatment effect is that strong that even the worst-comparison analysis cannot reduce it to a not significant level.
For that reason a subgroup was used in this thesis which had a smaller treatment effect. The tipping point analysis ended successfully at $\delta = 4.1$. This implies that the hazard ratio of patients staying in the active treatment group vs. censored patients is $1/4.1 = 0.24$ which is unlikely. In conclusion, even a more marginally-significant subgroup of the provided study with smaller treatment effect can be judged robust for informative censoring when using the tipping point method as sensitivity analysis.

One option that has not been examined in this thesis is a delta adjustment with a $\delta < 1$. Such a parameter value for δ would signify that patients who drop out of a study are assumed to have later event times than uncensored patients. This can be appropriate for a trial where patients in the reference group may receive rescue medication. If that is the case they will do better after the drop out because of it. The imputation could be done by means of the KM curve for the reference group and a $\delta < 1$ to simulate this. The reason for ignoring this option is that this thesis focused on sensitivity analyses where it is typically not assumed that censored patients actually do better after drop out. If one wanted to apply it has to be handled carefully. The assumption of an improvement after drop out needs strong evidence.

The second approach is based on an idea presented by Roger, a reference-based imputation for longitudinal data [19][20]. This idea was transferred to time-to-event data, which means that patients who drop out of the active treatment group behave like the patients of the reference group regarding the event time.
In contrast to the tipping point analysis, the reference-based imputation does not "search" for a scenario where the treatment effect is not significant any more. If the treatment effect is still significant after the censorings have been imputed, the analysis can be considered robust.
Since in most oncology studies the reference group receives the best licensed treatment the assumption of the reference based imputation are reasonable and applicable in the real world. However, one has to be careful using it in studies where progression is an endpoint. Patients leaving a study are often in worse health than patients remaining in the study. It may be that those patients would do even worse than the patients in the reference group. In this case the actual hazard of censored patients after drop out would be even higher than

assumed in the reference-based analysis. The pattern imputation method can deal with such assumptions but not the reference-based analysis. In case it is assumed that patients who drop out actually have the same hazard than patients in the reference group, the reference based imputation can be considered as sensitivity analysis.
Since the δ in the delta adjustment can be interpreted as hazard after the drop out for censored patients, the reciprocal of the hazard ratio of the original study could be applied to the active treatment group to imply that the patients of the active treatment group have the same hazard like the patients in the reference group after drop out as suggested in Zhao et al. [35]. The study in this thesis hat a hazard ratio of 0.57 comparing active treatment group vs. reference group which would mean $\delta = 1.75$ for the active treatment group after censoring. Interestingly, the δ closest to the results of the reference-based imputation in this thesis is 1.57. This difference can potentially be explained by the fact that the reference-based imputation uses a different distribution (of the reference group) for the imputation than the active treatment distribution used in Zhao et al.
The reference-based imputation method led to a hazard ratio of 0.599 comparing active treatment vs reference which is slightly closer to unity than the original study analysis (HR 0.57). But the treatment effect is really strong, even after the imputation which is also shown with a corresponding p-value of 0.0004. The original study analysis therefore can be judged robust to informative censoring after applying a reference-based imputation as sensitivity analysis.

It is unlikely that all patients who drop out of a study meet the same expectations with the progress of their disease after they left the trial. Based on this, the pattern imputation approach was developed and implemented. The pattern imputation can be used not only as a sensitivity analysis but also to get an impression of how the results might look like if the data would not have been censored.
The big advantage of the pattern imputation is that different assumptions can be made for theoretically every censoring. One has to be careful with the number of censorings in the patterns. If there are too much patterns with a small amount of censorings the outcome of the imputations can be imprecise. This is because different distributions are defined for each treatment in every pattern. If the imputation procedure produces extreme values for the imputation time it influences the results much more if there is a small amount of censorings in this pattern. The combination of imputed observations from different distributions can potentially cause imprecision. This is due to less censorings being imputed with more different distributions and therefore the patterns have a higher variance if they are smaller. This issue can get even more serious if the imputation is planned to be done separated by a stratification variable. As mentioned in the previous approaches a stratified imputation decreases the number of events used to calculate the respective Kaplan Meier curves which can influence the results if the number of events is not big enough.
If there are enough events and enough censorings in each pattern the pattern imputation is a very powerful tool not only for sensitivity analyses. The usage of the implementation of the pattern imputation is very easy with the instruction data set.
One problem with the pattern imputation that has not been addressed in this thesis is the construction of the patterns and the identification of patients belonging to this pattern. The study provided for this thesis included the reasons for drop out that could be used to form patterns. A manual of how to find patterns of censorings in time-to-event data could be useful additionally to the imputation program.
In this thesis 3 examples for pattern imputation are given. None of them ended up making the treatment effect not significant with example 1 being closest to it by switching the Kaplan Meier curves for the KMI for the treatments. Example 3 for the pattern imputation is the most complex with having 4 patterns with different assumptions. But is is also the most realistic in points of behaviour after censoring. The hazard ratio of active treatment vs. reference treatment is slightly bigger than in the original study results (0.6 after imputation compared to 0.57 in the original study) but it is not even close to being not significant (p-value=0.0004). This

example also shows the strength of the pattern imputation. The tipping point analysis and the reference-based analysis both only give one option for adjusting the Kaplan Meier curve of the active treatment group. The pattern imputation can handle different assumptions not only for the different treatment groups but also for pre specified patterns of censorings in one treatment group or even stratum if there are enough observations.

The implementation for the pattern imputation is done with two macros different from those for the first two approaches. This is due to the working process in this thesis which was started with the tipping point analysis as first extension to the CAR program because the implementation is the most straight forward. In fact, the delta adjustment and the reference-based analysis can be done with only the macros for the pattern imputation. For the first two approaches only one pattern is specified including all censorings where the reference treatment is imputed under the CAR assumption and the active treatment group is imputed under the respective assumptions for delta adjustment or reference-based analysis. To get the tipping point analysis working with the pattern imputation macros the implementation of the macro that increases the δ after every loop would have to be reworked.

In conclusion, all implemented methods work well and can be used under the right conditions. The assumptions for the approaches are realistic under the right conditions. For the tipping point analysis the boundaries should be chosen reasonable. The first two approaches suffer a little under the fact that only the active treatment curve can be adjusted with only one assumption but nonetheless they are very useful as sensitivity analysis tools and are relatively simple. Finally, it is noteworthy that the pattern imputation programs provide a method that has not been introduced before for time to event data which makes it particularly interesting.

List of Figures

List of Tables

Bibliography

[1] David Collett. *Modelling survival data in medical research*. Chapman &Hall, 1st edition, 1995. ISBN 0-412-44880-7.

[2] David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society*, 34(2):187–220, 1972.

[3] Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[4] Major Greenwood. *A Report on the Natural Duration of Cancer*. Reports on public health and medical subjects. H.M. Stationery Office, 1926.

[5] Frank E. Jr. Harrel. *Regression Modeling Strategies*. Springer series in Statistics, 1st edition, 2001. ISBN 0-387-95232-2.

[6] Daniel F. Heitjan. Ignorability in general incomplete-data models. *Biometrika*, 81(4):pp. 701–708, 1994.

[7] John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Interscience, 2nd edition, 2002. ISBN 0-471-36357-X.

[8] Roderick J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):pp. 125–134, 1993. ISSN 01621459. URL `http://www.jstor.org/stable/2290705`.

[9] Roderick J. A. Little. A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3): 471–483, 1994. doi: 10.1093/biomet/81.3.471. URL `http://biomet.oxfordjournals.org/content/81/3/471.abstract`.

[10] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 1987. ISBN 0-471-80254-9.

[11] Roderick J.A. Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121, 1995.

[12] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3):163–70, 1966.

[13] Geert Molenberghs and Micheal G. Kenward. *Missing Data in Clinical Studies*. John Wiley & Sons, 2007. ISBN 978-0-470-84981-1.

[14] Michael O'Kelly and Ilya Lipkovich. Using multiple imputation and delta adjustment to implement sensitivity analyses for time-to-event data. *PSI Conference presentation*, 2014.

[15] Mahesh K. B. Parmar, Valter Torri, and Lesley Stewart. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in Medicine*, 17(24):2815–2834, 1998. ISSN 1097-0258.

[16] Richard Peto and Julian Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society*, 135(2):185–207, 1972.

[17] Bohdana Ratitch, Ilya Lipkovic, and Micheal O'Kelly. Combining analysis results from multiply imputed categorial data. PharmaSUG, 2013.

[18] Bohdana Ratitch, Michael O'Kelly, and Robert Tosiello. Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical Statistics*, 12(6):337–347, 2013. ISSN 1539-1612. doi: 10.1002/pst.1549. URL http://dx.doi.org/10.1002/pst.1549.

[19] James Roger and Mouna Akacha. Reference-based imputation for partially observed count data due to early withdrawal. *PSI conference slides*, 2014.

[20] James Roger and Michael O'Kelly. When and how to use reference based imputation for missing data. *PSI conference slides*, 2013.

[21] Mark D. Rothmann, Kallappa Koti, Kyung Yul Lee, Hong Laura Lu, and Li Yuan Shen. Missing data in biologic oncology products. *Journal of Biopharmaceutical Statistics*, 19:1074–1084, 2009.

[22] Donald B. Rubin. Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. *U.S. Department of Commerce*, pages 1–23, 1978.

[23] Donald B. Rubin. *Multiple Imputation for Nonresponse in surveys*. John Wiley & Sons, 1987. ISBN 0-471-08705-X.

[24] Donald B. Rubin and Nathaniel Schenker. Multiple imputation in health-are databases: An overview and some applications. *Statistics in Medicine*, 10(4):585–598, 1991. ISSN 1097-0258. doi: 10.1002/sim.4780100410. URL http://dx.doi.org/10.1002/sim.4780100410.

[25] Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[26] Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

[27] Lecia V. Sequist, James Chih-Hsin Yang, Nobuyuki Yamamoto, Kenneth O'Byrne, Vera Hirsh, Tony Mok, Sarayut Lucien Geater, Sergey Orlov, Chun-Ming Tsai, Michael Boyer, Wu-Chou Su, Jaafar Bennouna, Terufumi Kato, Vera Gorbunova, Ki Hyeong Lee, Riyaz Shah, Dan Massey, Victoria Zazulina, Mehdi Shahidi, and Martin Schuler. Phase iii study of afatinib or cisplatin plus pemetrexed in patients with metastatic lung adenocarcinoma with egfr mutations. *Journal of Clinical Onclogy*, 31(27):3326–3337, September 2013.

[28] Fotios Siannis, John Copas, and Guobing Lu. Sensitivity analysis for informative cenosring in parametric survival models. *Biostatistics*, 6(1):77–91, 2005.

[29] Robert E. Tarone and James Ware. On distribution-free tests for equality of survival distributions. *Biometrika*, 64(1):156–160, 1977. doi: 10.1093/biomet/64.1.156. URL http://biomet.oxfordjournals.org/content/64/1/156.abstract.

[30] Jeremy M.G. Taylor, Susan Murray, and Chiu-Hsieh Hsu. Survival estimation and testing via multiple imputation. *Statistics & Probability Letters*, 58(3):221 – 232, 2002. ISSN 0167-7152. doi: http://dx.doi.org/10.1016/S0167-7152(02)00030-5. URL http://www.sciencedirect.com/science/article/pii/S0167715202000305.

[31] Anastasios A. Tsiatis and Marie Davidian. Multiple imputation methods for testing treatment differences in survival distributions with missing cause of failure. *Biometrika*, 89(1):238, 244 2002.

[32] Stefan Wellek. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC, 2nd edition, 2010.

[33] Margaret C. Wu and Kent Bailey. Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine*, 7(1-2):337–346, 1988. ISSN 1097-0258. doi: 10.1002/sim.4780070134. URL `http://dx.doi.org/10.1002/sim.4780070134`.

[34] Yang C. Yuan. Multiple imputation for missing data: Concepts and new development(version 9.0). Technical report, SAS Institute Inc. URL `http://support.sas.com/rnd/app/stat/papers/multipleimputation.pdf`.

[35] Yue Zhao, Amy H. Herring, Haibo Zhou, Mirza W. Ali, and Gary W. Koch. A multiple imputation method for sensitivity analyses of time-to-event data with possibly informative censoring. *Journal of Biopharmaceutical Statistics*, 24(2):229–253, March 2014.

Name: Simon Fink

Statutory Declaration

I declare that I have developed and written this master's thesis completely by myself, and have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked. This master's thesis was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere.

Munich, 06.05.2015 .....

Simon Fink